Welcome To

ISSCC 2014
Session 19
Nonvolatile Memory Solutions

A 128Gb MLC NAND-Flash Device Using 16nm Planar Cell

Mark Helm[1], Jae-Kwan Park[1], Ali Ghalam[1], Jason Guo[1], Chang wan Ha[1], Cairong Hu[1], Heonwook Kim[1], Kalyan Kavalipurapu[1], Eric Lee[1], Ali Mohammadzadeh[1], Dan Nguyen[1], Vipul Patel[1], Ted Pekny[1], Bill Saiki[1], Daesik Song[1], Jeff Tsai[1], Vimon Viajedor[1], Luyen Vu[1], Tinwai Wong[1], Jung Hee Yun[1], Ramin Ghodsi[1], Andrea D'Alessandro[2], Domenico Di Cicco[2], Violante Moschiano[2]

[1]Micron Technology, San Jose, CA

[2]Micron Technology, Avezzano, Italy

Outline

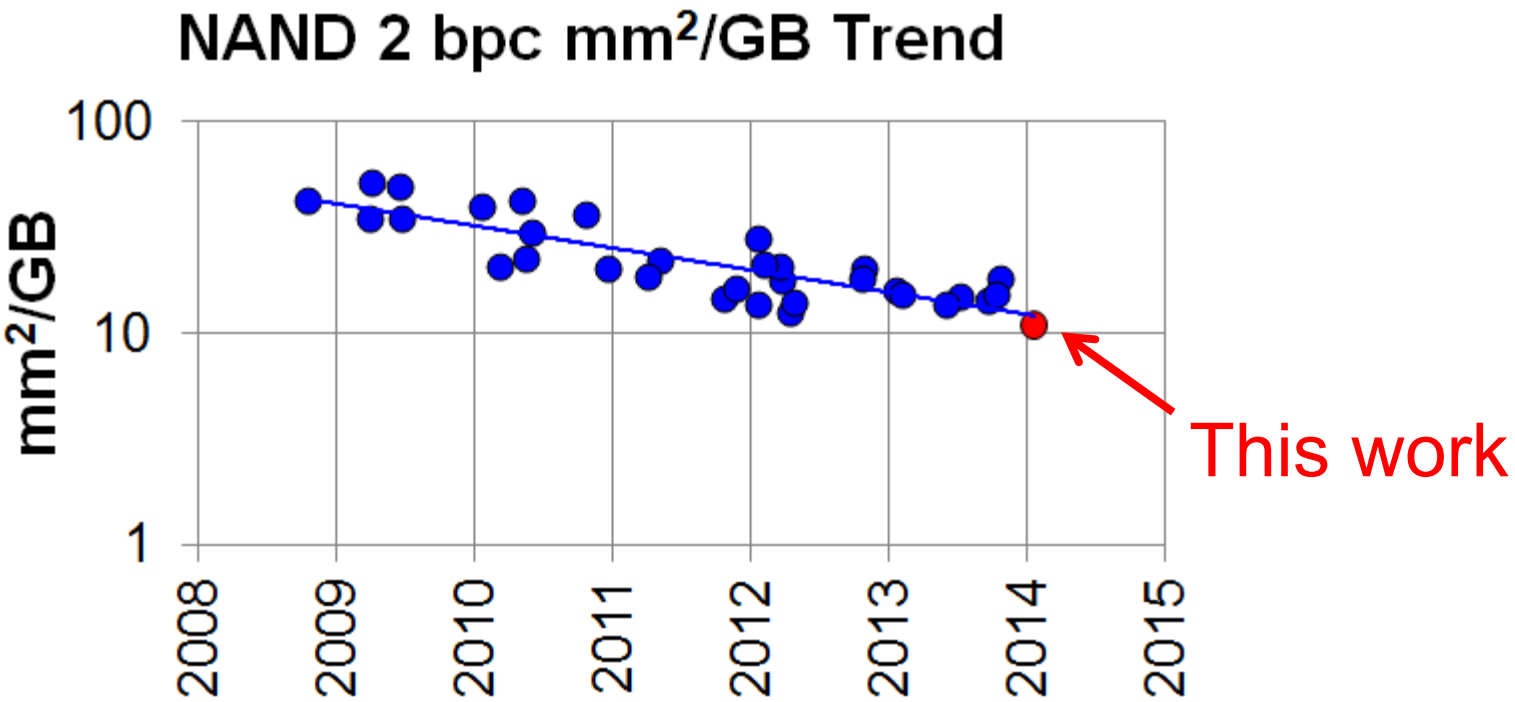
- Introduction
- Technology and Device Metrics
- Lower Page Pre-read Error Mitigation
- Boosted Bitline Negative Sense
- User-selectable Power Management
- Conclusion

Outline

- Introduction
- Technology and Device Metrics
- Lower Page Pre-read Error Mitigation
- Boosted Bitline Negative Sense
- User-selectable Power Management
- Conclusion

Introduction

- NAND memory scaling continues to maintain an exponential pace
- 16nm technology node extends this trend

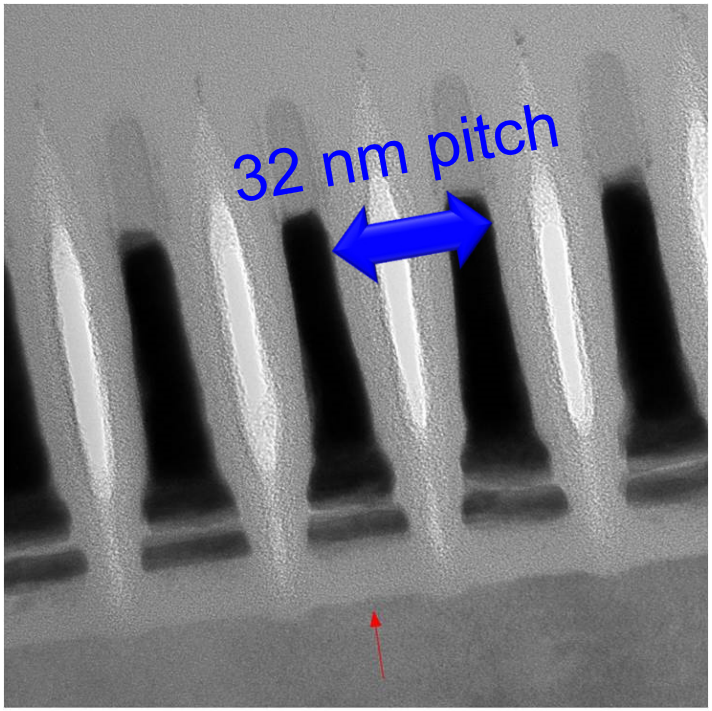


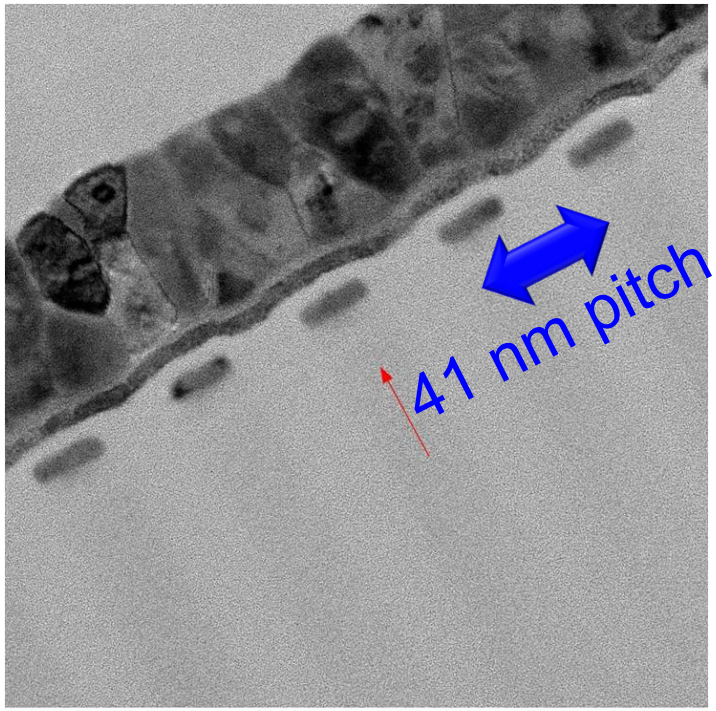
Outline

- Introduction
- Technology and Design Attributes
- Lower Page Pre-read Error Mitigation
- Boosted Bitline Negative Sense
- User-selectable Power Management
- Conclusion

16nm Planar Cell Technology

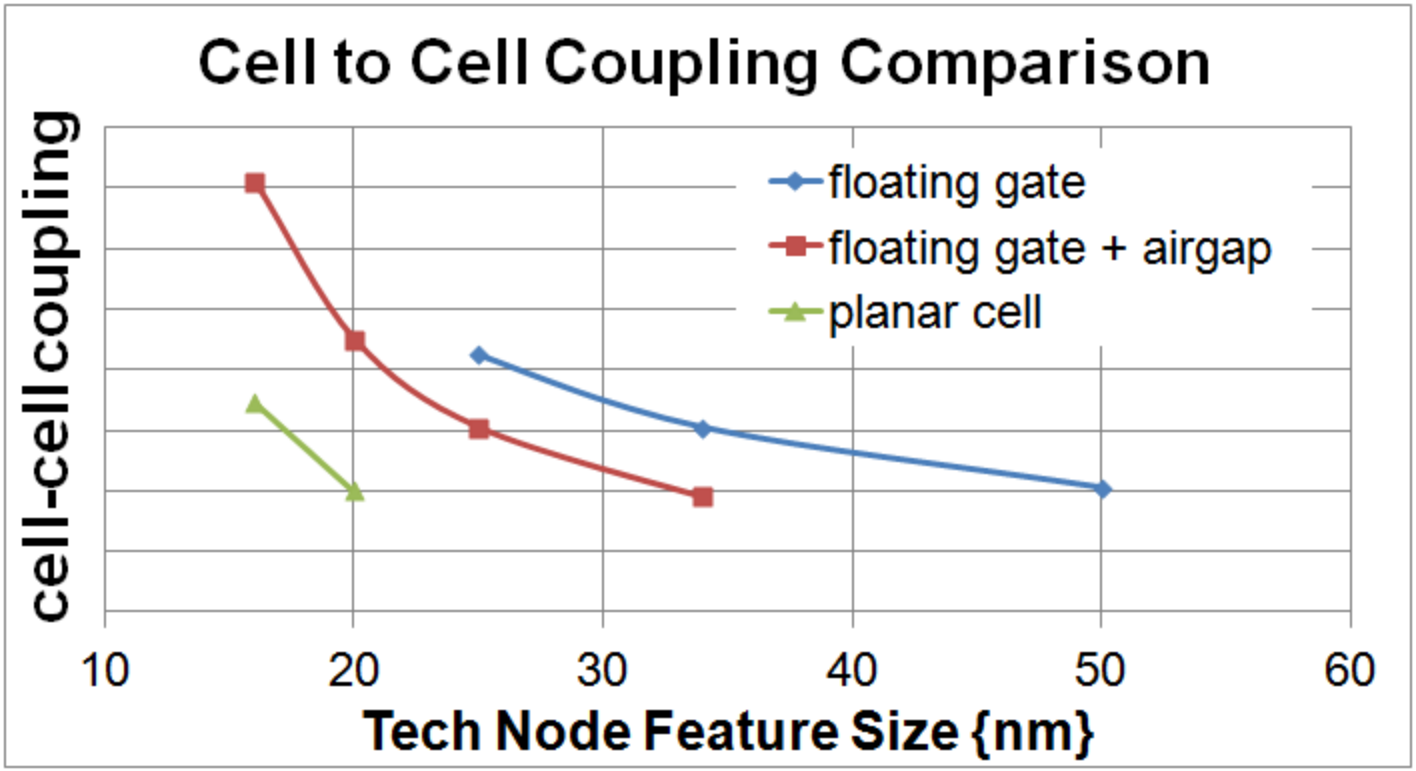
- 2nd Generation Planar Cell
- Optimized hi-k dielectric/metal gate
- Wordline airgap





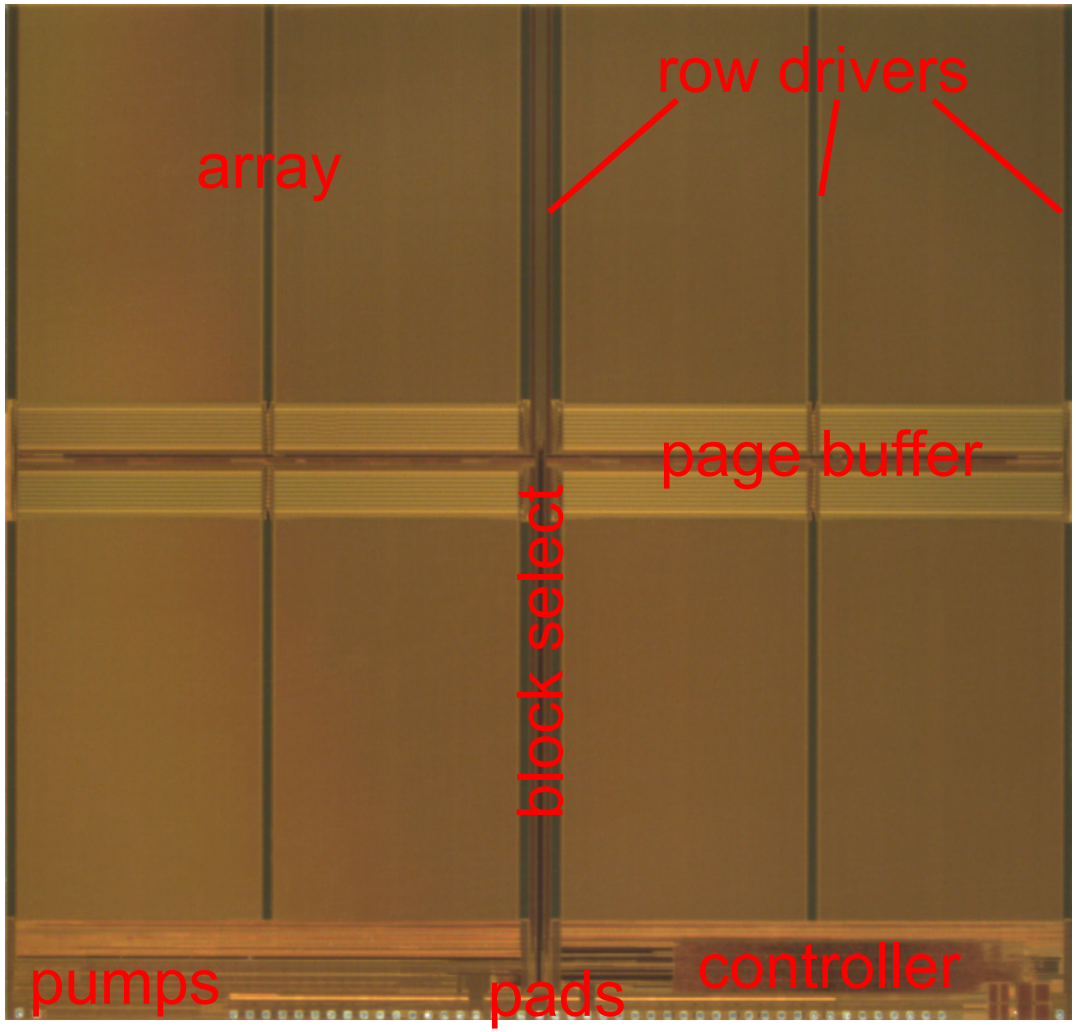
Cell Coupling Improvement

- Planar cell reduces cell to cell coupling
- Hi-k dielectric maintains adequate gate coupling ratio



128Gb Die Architecture

- Two planes
- 16kB page
- Shielded bitline
- Center page buffer architecture



128Gb Die Attributes

- Optimized for client and enterprise SSD adoption
  - Fast read and write access times
  - High performance, low power I/O

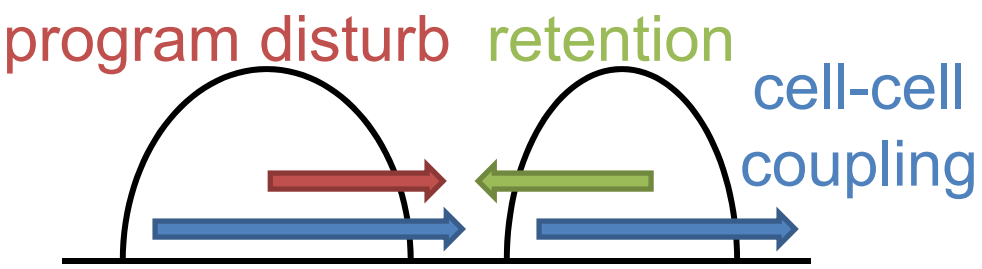
Technology	16nm planar cell 128 wordline string 3 metals
Density	128Gb
Bits per cell	2
Architecture	2 plane 16kB page/plane Shielded bitline Center page buffer
Die Size	173.3mm^2
Average read time	45µs
Average pgm time	1185µs
Average erase time	3ms
I/O	400 MT/s/pin ONFI 3
Power Supply	2.7 – 3.6V Vcc 1.70 – 1.95V Vccq

Outline

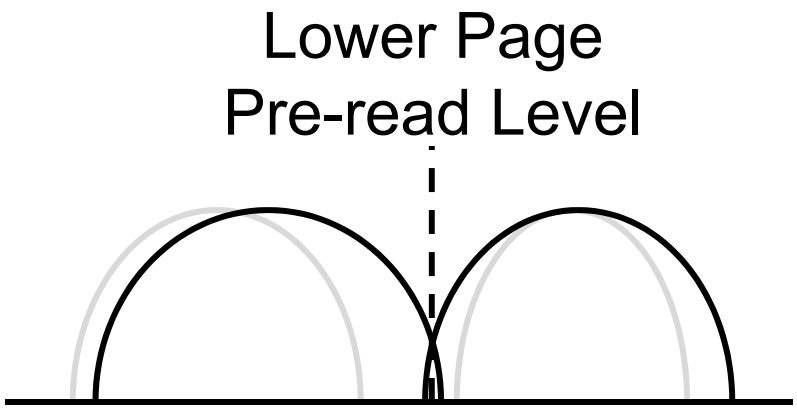
- Introduction
- Technology and Design Attributes
- Lower Page Pre-read Error Mitigation
- Boosted Bitline Negative Sense
- User-selectable Power Management
- Conclusion

Lower Page Vt Placement

- After lower page placement, broadening of the Vt distribution occurs
  - Cell to cell coupling
  - Program disturb
  - Retention charge loss



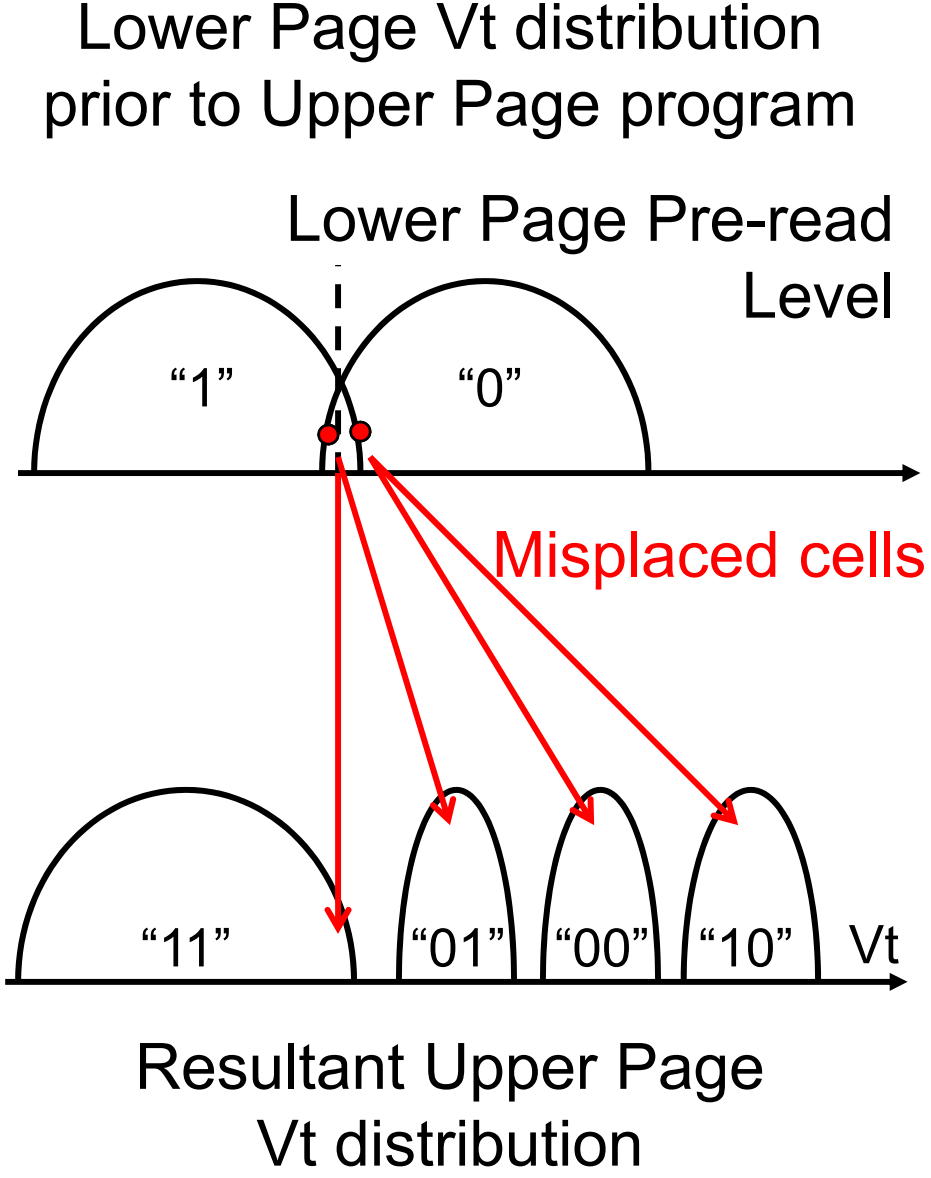
As-placed Lower Page Vt distribution



Lower Page Vt distribution prior to Upper Page program

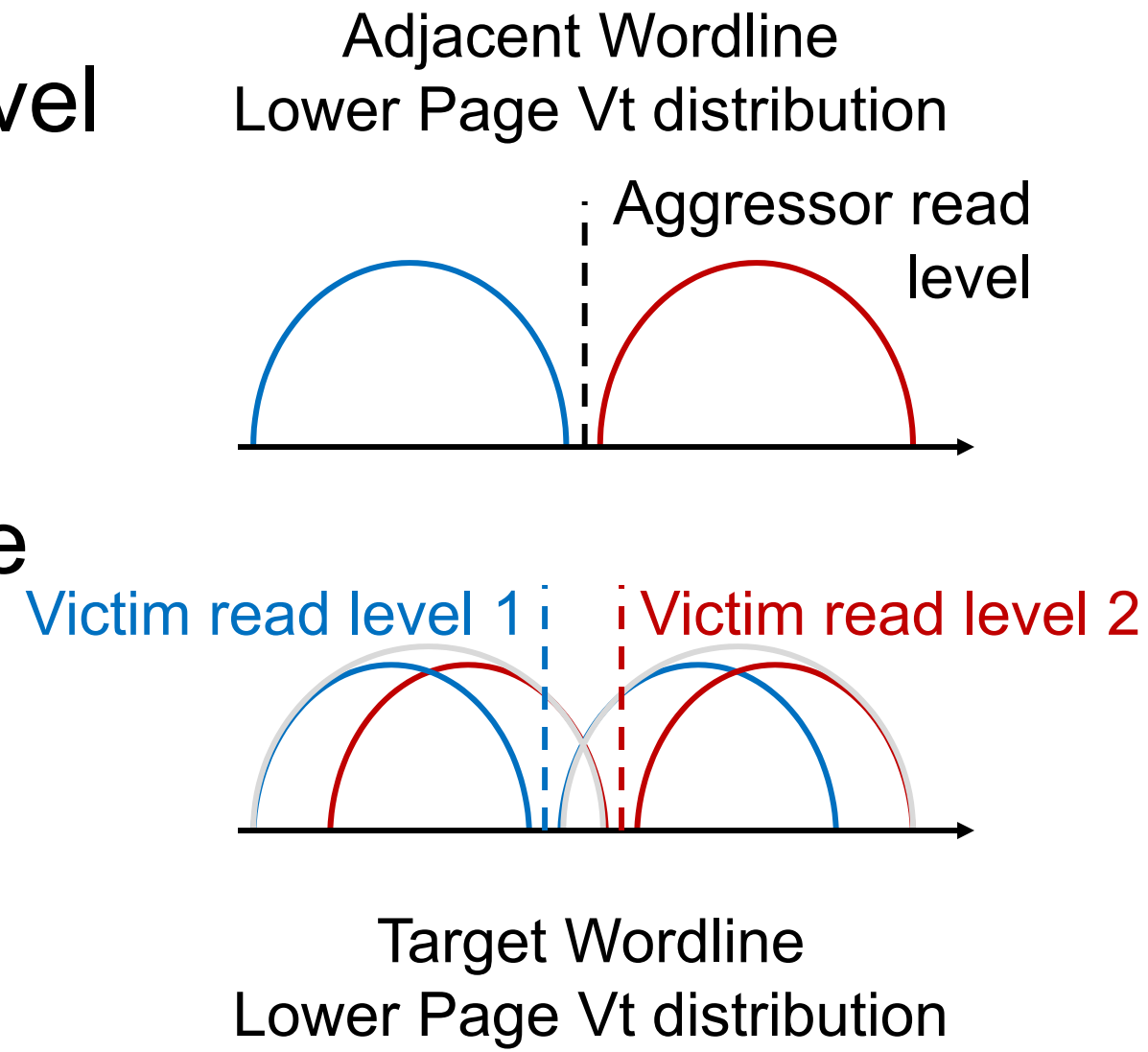
Upper Page Program Misplacement

- Pre-read errors lead to misplacement during upper page program
- Misplacement errors highly undesirable for advanced ECC algorithms



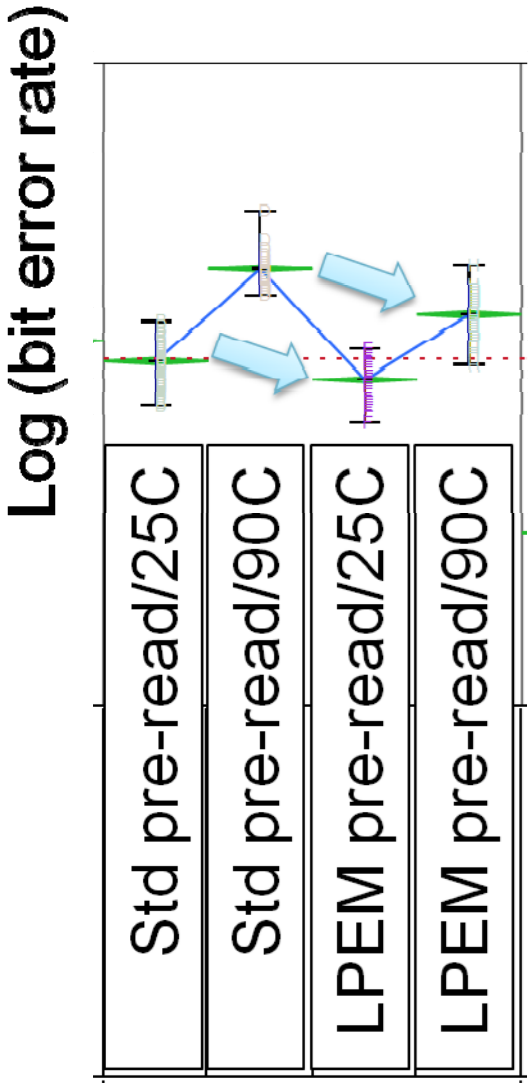
Corrective Lower Page Pre-Read

- Adjust read level of victim cell based on aggressor cell state to reduce errors



Misplacement Reduction

- Bit error rate effectively reduced by deploying LPEM feature
- Advanced ECC efficacy improved

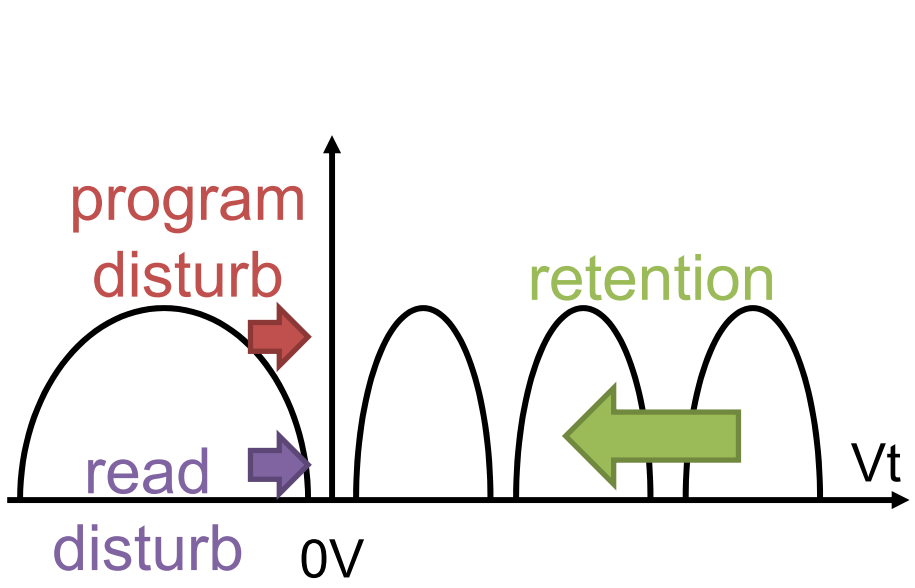


Outline

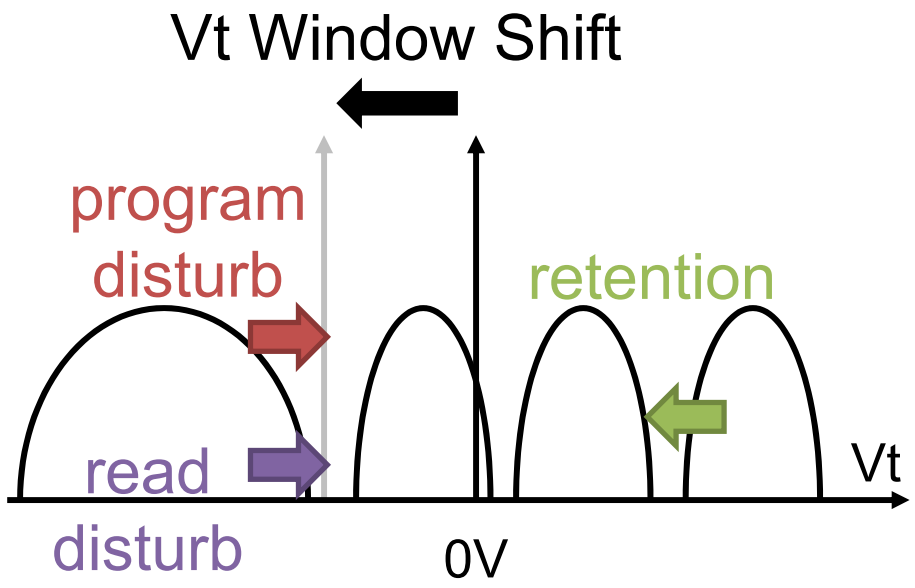
- Introduction
- Technology and Design Attributes
- Lower Page Pre-read Error Mitigation
- Boosted Bitline Negative Sense
- User-selectable Power Management
- Conclusion

Vt Window Optimization

- Ability to shift the Vt window is critical
- Balance program and read disturb versus retention charge loss



Unbalanced Vt Window

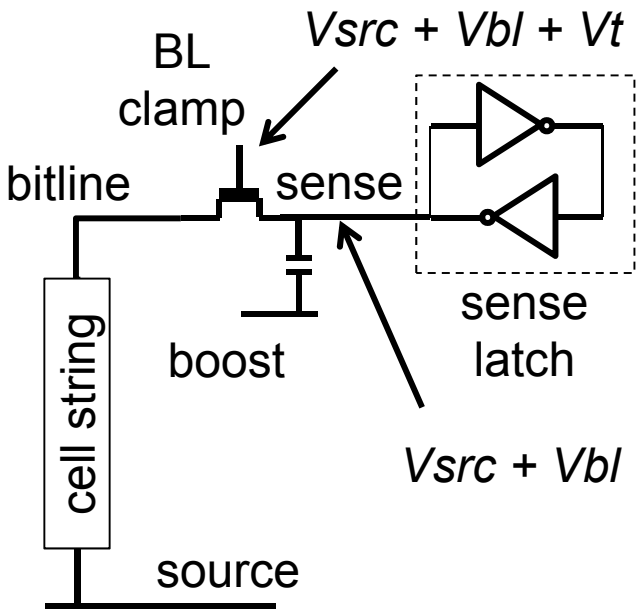


Balanced Vt Window

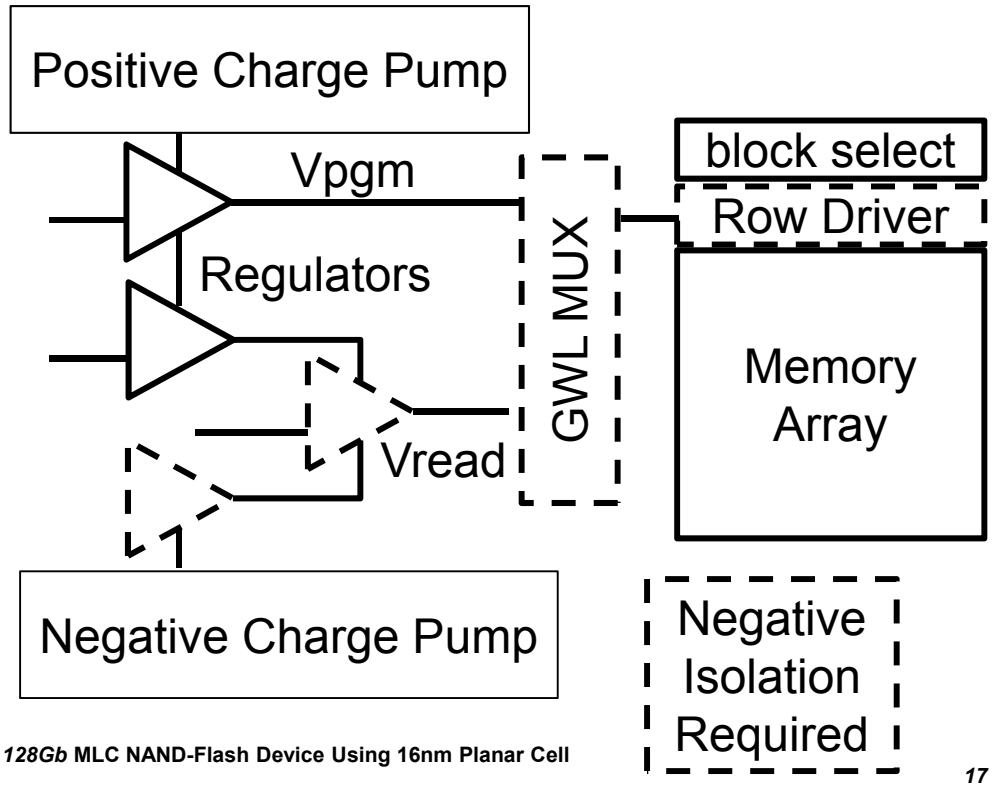
Vt Window Optimization

- Negative Vt window shift options are limited or costly

Source Bias Negative Window Shift

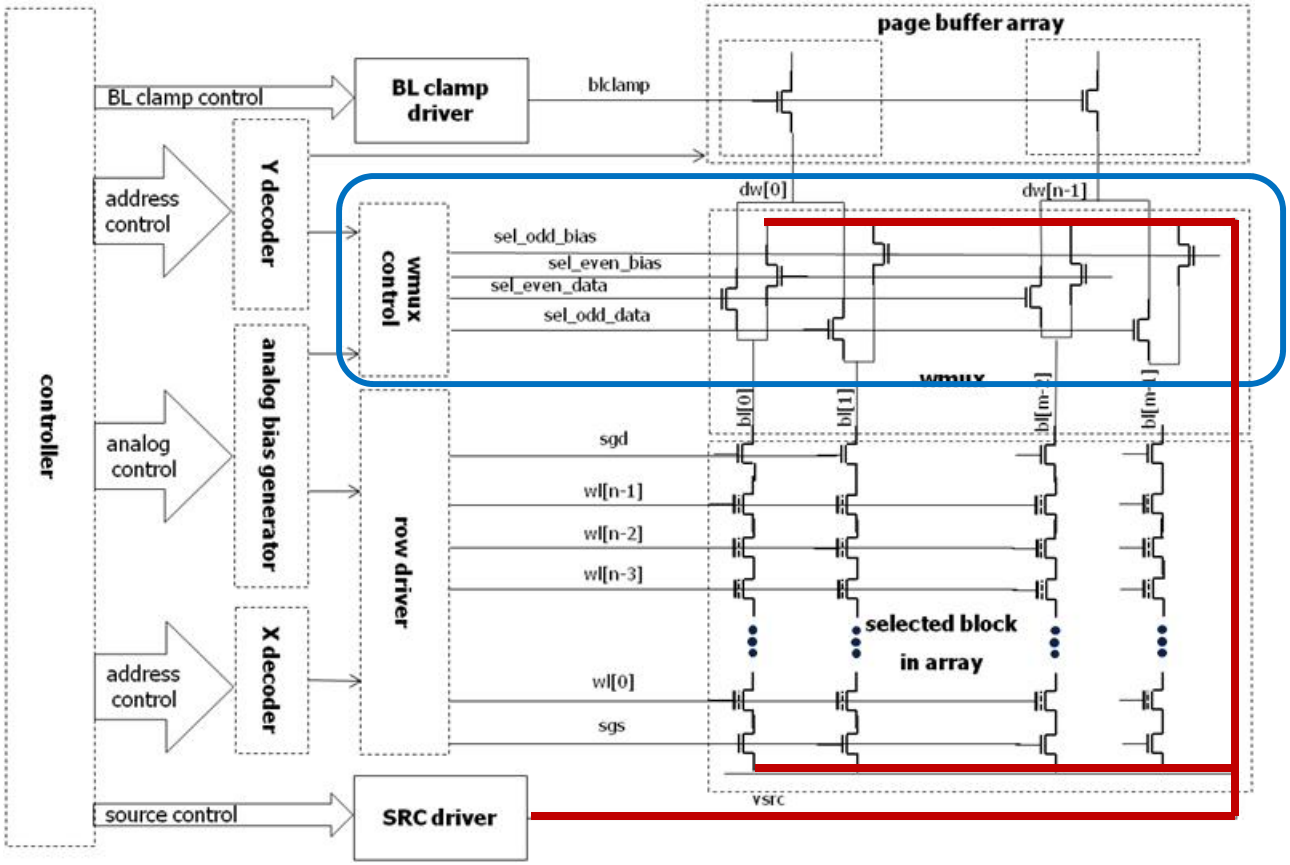


Negative Wordline Bias



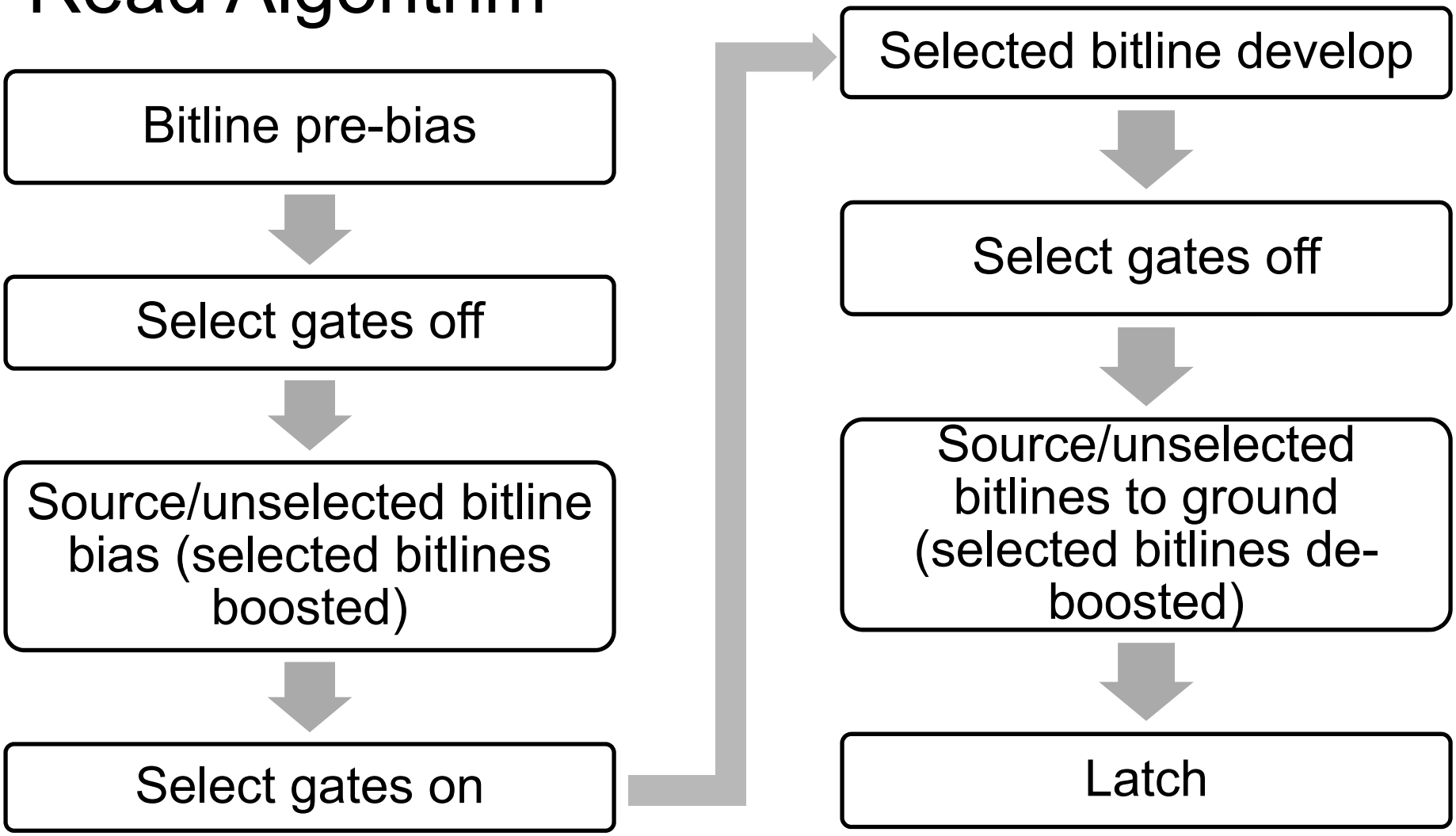
Bitline Circuit Architecture

- Selected bitline connected to the page buffer
- Unselected bitlines connected to source via bitline multiplexer circuit



Boosted Bitline Negative Sense

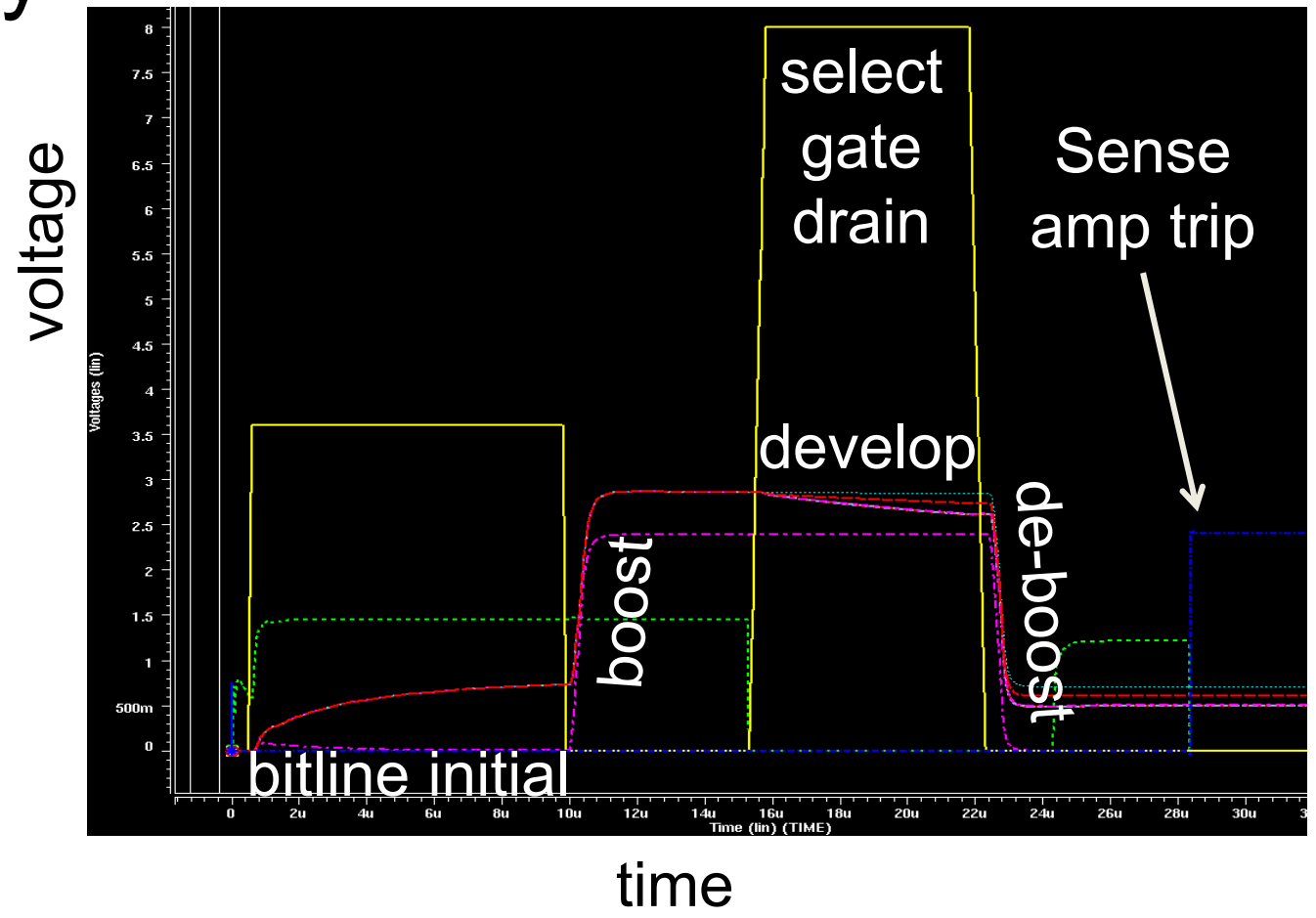
- Read Algorithm



Boosted Bitline Negative Sense

- Waveform shows critical signals for functionality

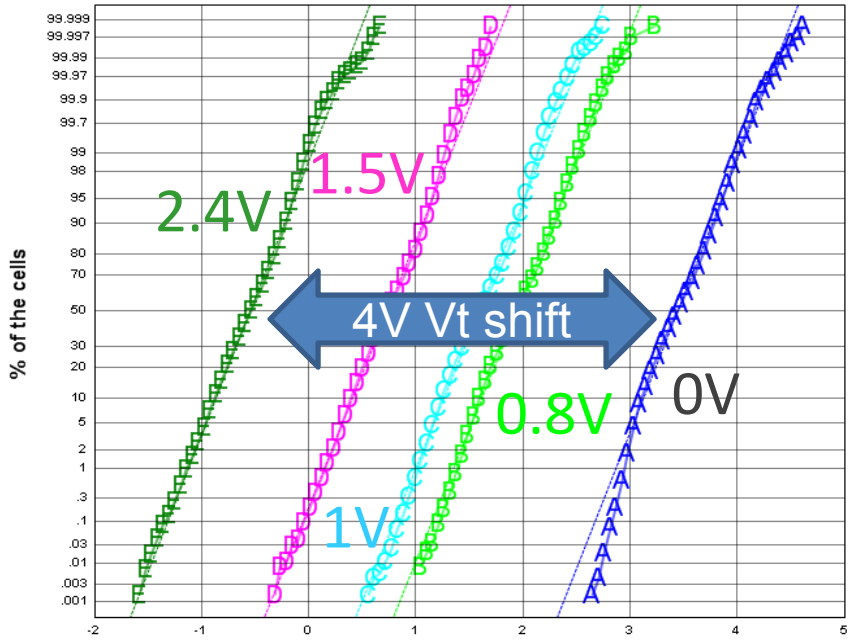
Waveform for Boosted Bitline Negative Sense (BBNS)



Vt shift with BBNS

- Large Vt shift capability
- No degradation in Vt distribution width with Vt window shift

Sensed Cell Vt vs. Boost Voltage for BBNS



A Vboost = 0V

B Vboost = 0.8V

C Vboost = 1.0V

D Vboost = 1.5V

E Vboost = 2.4V

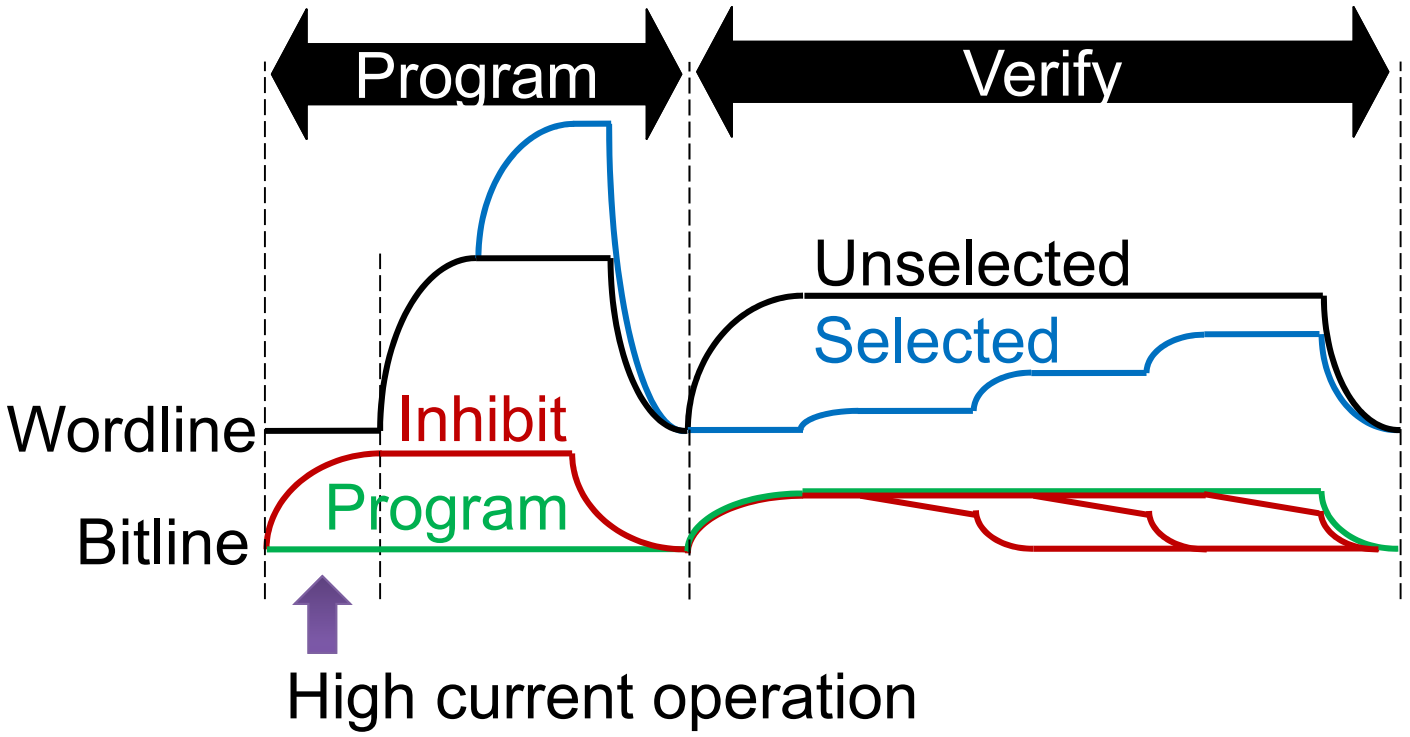
Vt {V}

Outline

- Introduction
- Technology and Design Attributes
- Lower Page Pre-read Error Mitigation
- Boosted Bitline Negative Sense
- User-selectable Power Management
- Conclusion

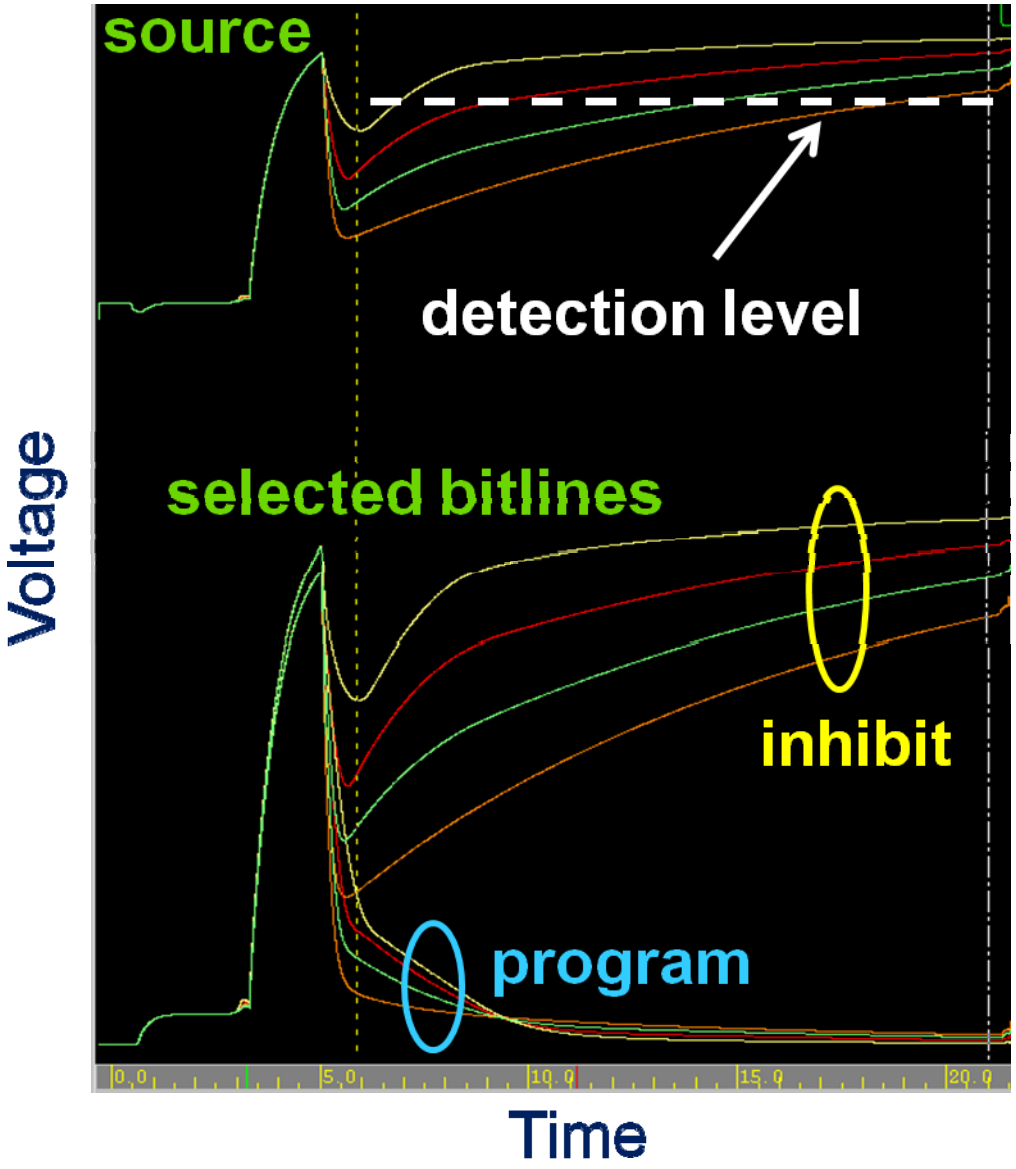
Program - Power Attributes

- Bitline bias phase during program is high power
- Power/performance envelope flexibility based on market need



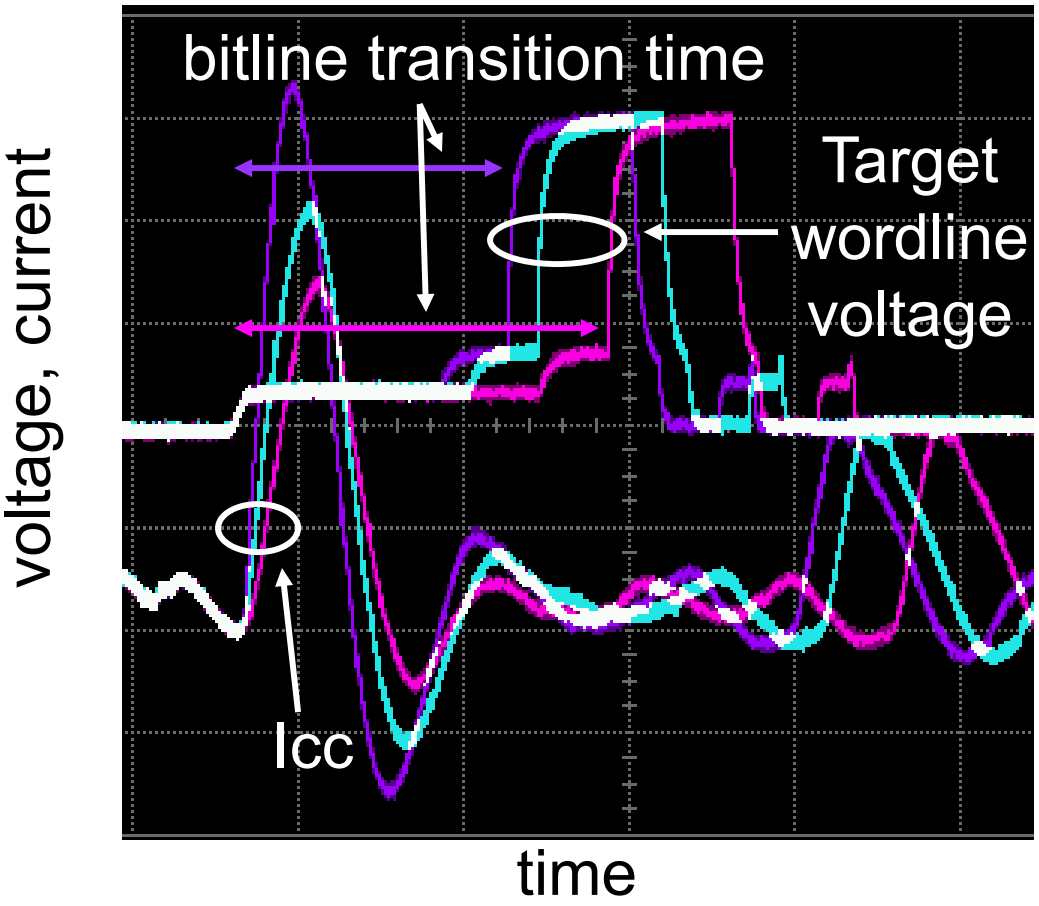
Bitline Bias Detection

- Peak power reduction limits current available for bitline bias
- Source driver is monitored with detector to signal completion of bitline bias phase



Results

- User-selectable peak power
- Automatic time adjustment for bitline biasing



Outline

- Introduction
- Technology and Design Attributes
- Lower Page Pre-read Error Mitigation
- Boosted Bitline Negative Sense
- User-selectable Power Management
- Conclusion

Conclusion

- 2nd generation hi-k/metal gate 16nm planar cell technology used to further reduce die size and cost for our 2 bpc 128Gb device.
- Design techniques employed:
  - Lower page error mitigation minimizes program misplacement errors improving compatibility with advanced ECC.
  - Boosted bitline negative sense allows further negative Vt window shift opportunities.
  - User selectable power management feature allows power – performance trade-off to be managed depending on market requirements.

A 93.4mm^2 64Gb MLC NAND-Flash Memory with 16nm CMOS Technology

Sungdae Choi, Duckju Kim, Sungwook Choi, Byungryul Kim, Sunghyun Jung, Kichang Chun, Namkyeong Kim, Wanseob Lee, Taisik Shin, Hyunjong Jin, Hyunchul Cho, Sunghoon Ahn, Yonghwan Hong, Ingon Yang, Byoungyoung Kim, Pilseon Yoo, Youngdon Jung, Jinwoo Lee, Jaehyeon Shin, Taeyun Kim, Kunwoo Park, Jinwoong Kim

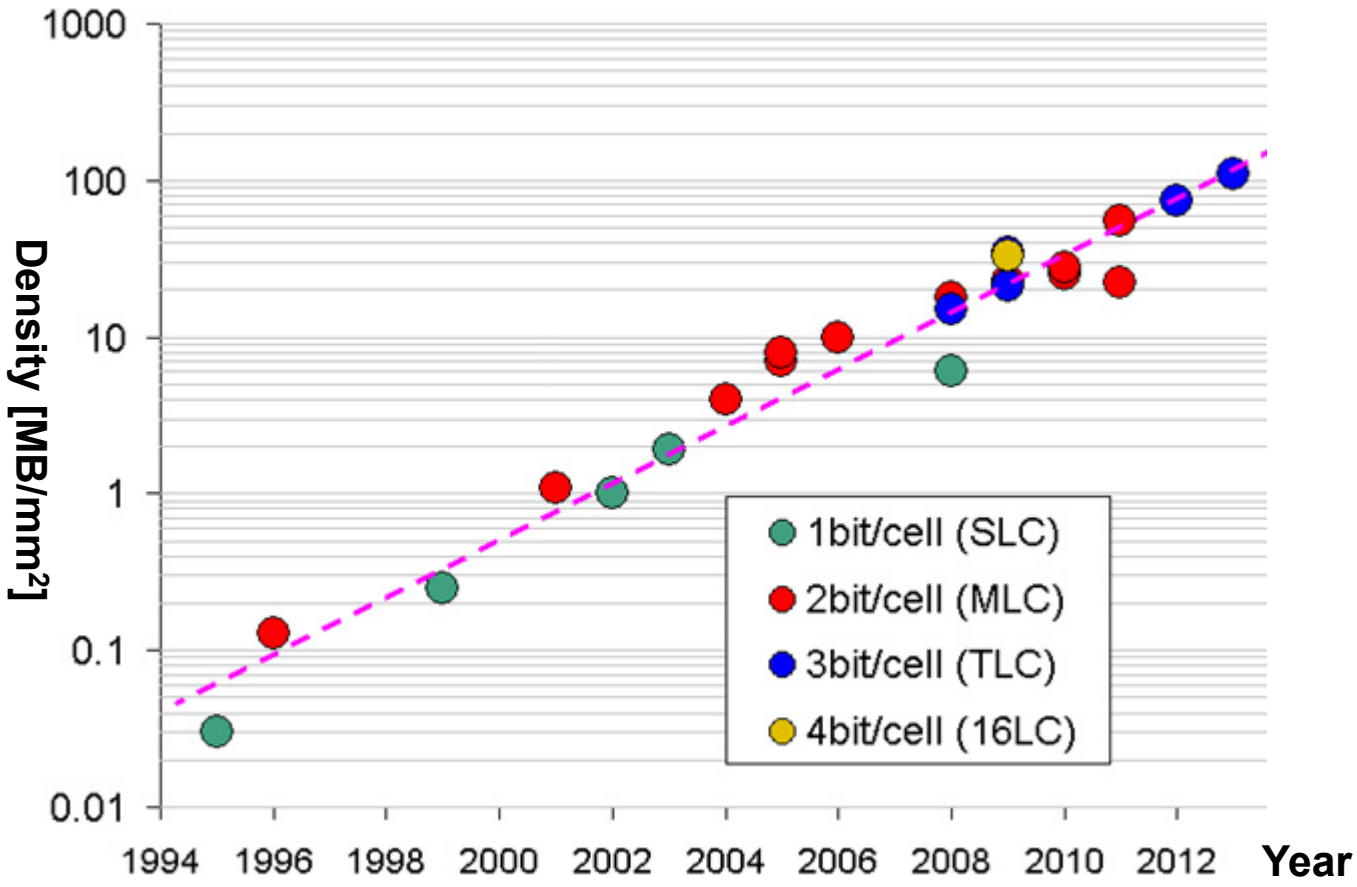
SK hynix, Icheon, Korea

Outline

- Introduction
  - Demand for high-density NAND flash
  - Hurdles for NAND Tech. shrink
- Proposed Schemes
  - Cell-current-controlled screen out sensing
  - Delayed P1 PGM Pulse
  - Peak current control
  - Load-balanced control signal P&R
- Summary

NAND Flash Density Trend

- Rising demand for high-density NAND flash
  - Tech. shrink for higher density & smaller package

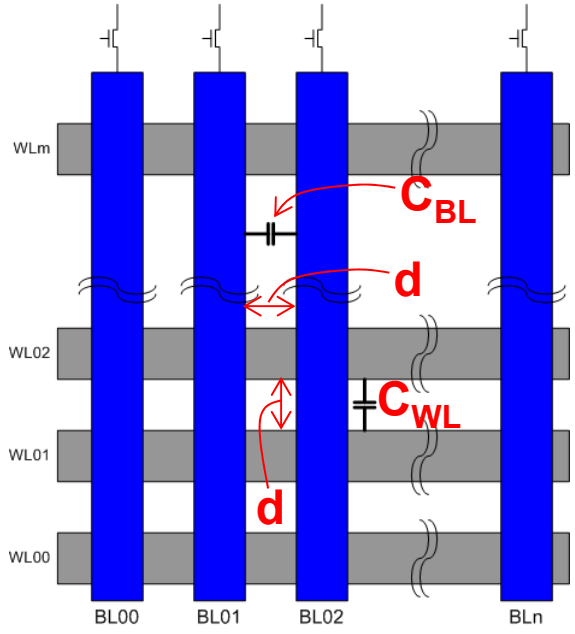


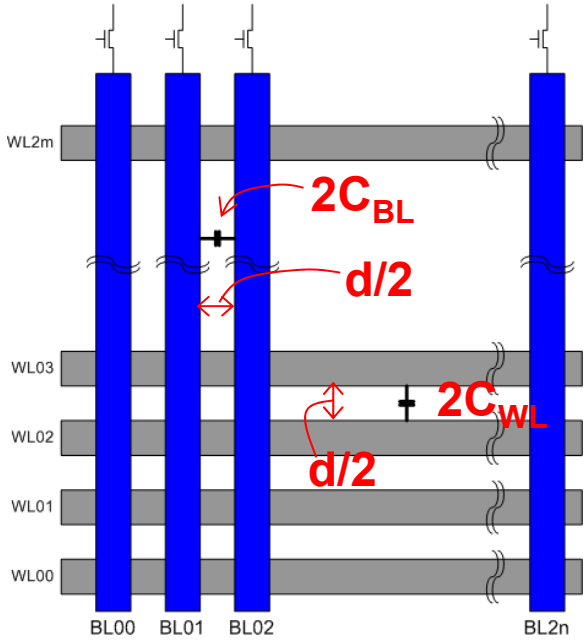
Hurdles for Tech. Shrink

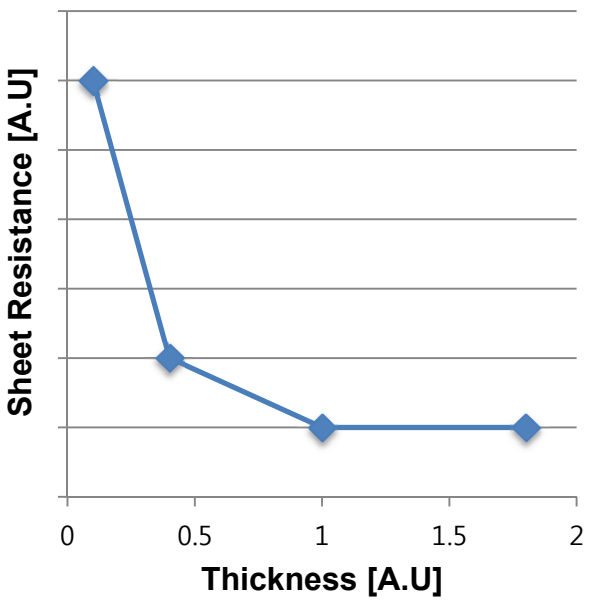
- Smaller cell transistor
  - Smaller cell current
  - Need of higher-sensitive sensing circuits
- Larger capacitive coupling
  - WL-to-WL and BL-to-BL
  - Larger cell distribution
  - Larger power consumption
- Larger R due to narrower metal width
  - Slower BL precharge time
  - Read/Write performance degradation

BL & WL Loading

- Tech. shrink increases RC loading.
  - Not chip size reduction, but density increase.
- Deep sub-micron tech. shows non-linear increase in sheet R.







Breakthrough

- 4 proposed schemes contribute reducing
  - (Peak) Power consumption
  - Cell distribution
  - Die size

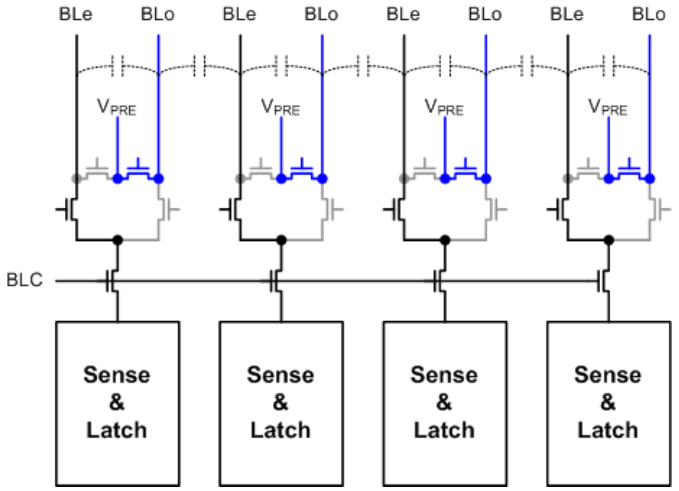
Target	Proposed Scheme
Smaller Power Consumption	• Peak-current control in BL controller
	• Cell-current-controlled screen-out sensing scheme
Narrow Cell Distribution	• Delayed P1 PGM Pulse
Smaller Die Size	• Load-balanced control signal block P&R

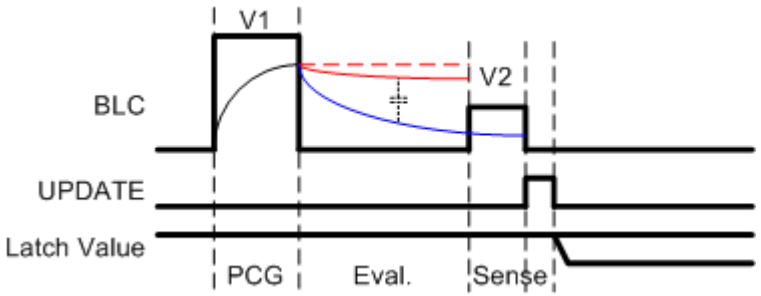
Outline

- Introduction
  - Demand for high-density NAND flash
  - Hurdles for NAND Tech. shrink
- Proposed Schemes
  - Cell-current-controlled screen out sensing
  - Delayed P1 PGM Pulse
  - Peak current control
  - Load-balanced control signal P&R
- Summary

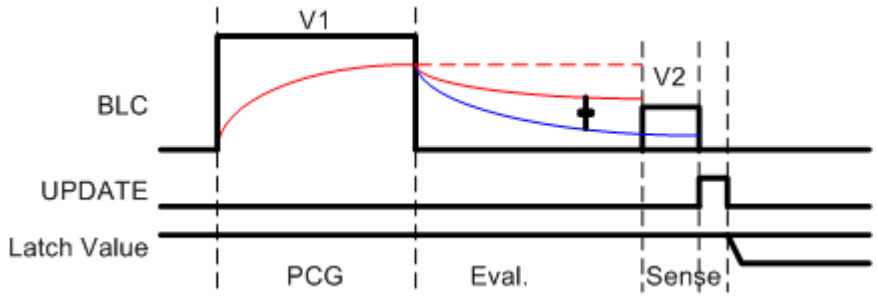
BL Shielding Structure

- BL shielding structure has disadvantage on deep sub-micron NAND flash memory.
  - ½ page is accessible at a time.
    - R/W performance degradation
  - Large BL-to-BL parasitic cap.
    - Longer precharge/evaluation time
    - Larger power to precharge BLs
    - Smaller off-cell sensing margin





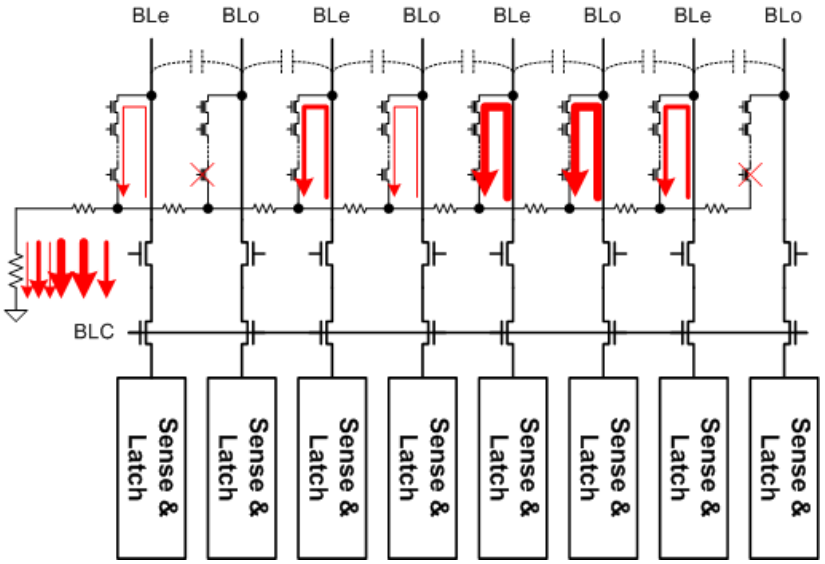
Sensing with small RC



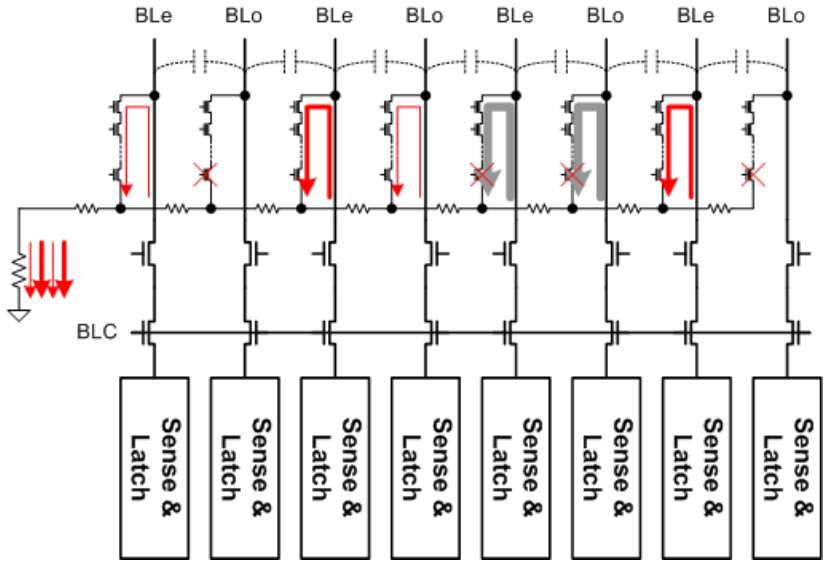
Sensing with large RC

All-Bit-Line Structure

- No BL-to-BL cap at BL precharge
  - All BLs have the same level.
- Large amount of current sink
  - Causes source-line level boosting → Degrades sensing accuracy
  - Large power consumption
- Screen-out operation filters large current cells
  - I_{cell} >> sensing criterion



ABL Sensing

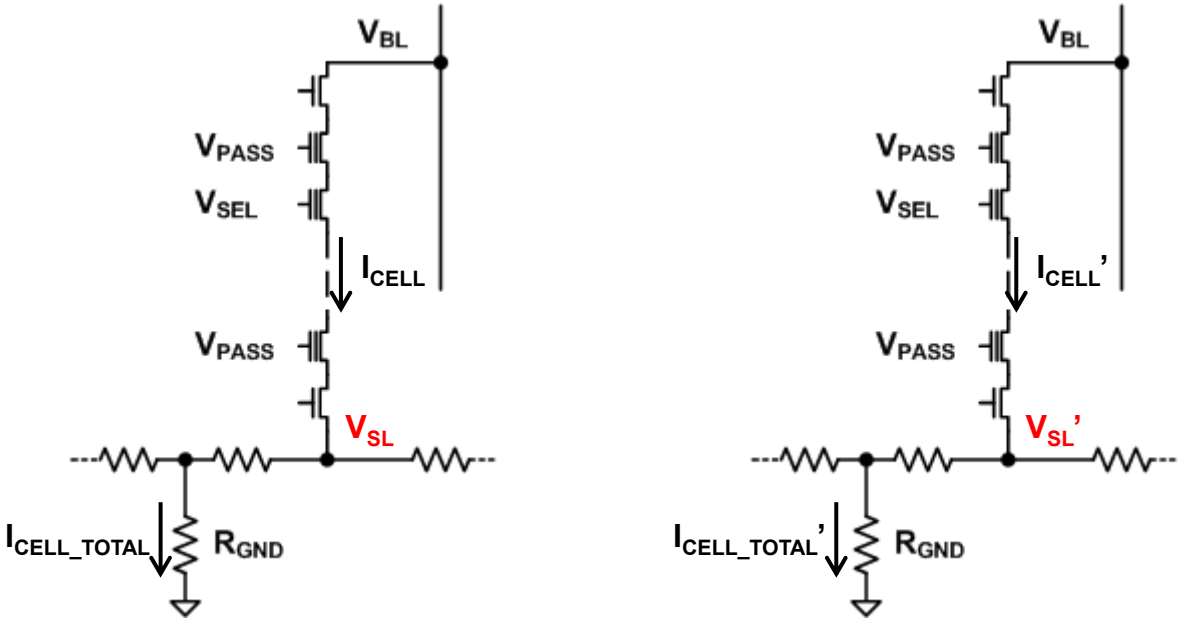


ABL with Screen-out operation

During Screen-out Operation

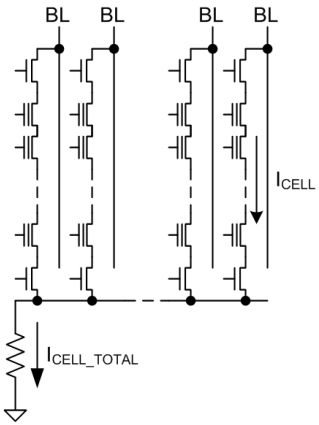
- Large-current cells still remain.
  - $I_{CELL} \propto f(V_{BL}-V_{SL})$
  - Inaccurate cell-current due to source-line level boosting
    - "Cell current" is defined value at V_{SL}=0V.
    - V_{SL} > 0V flows smaller cell current at the same V_{BL}.

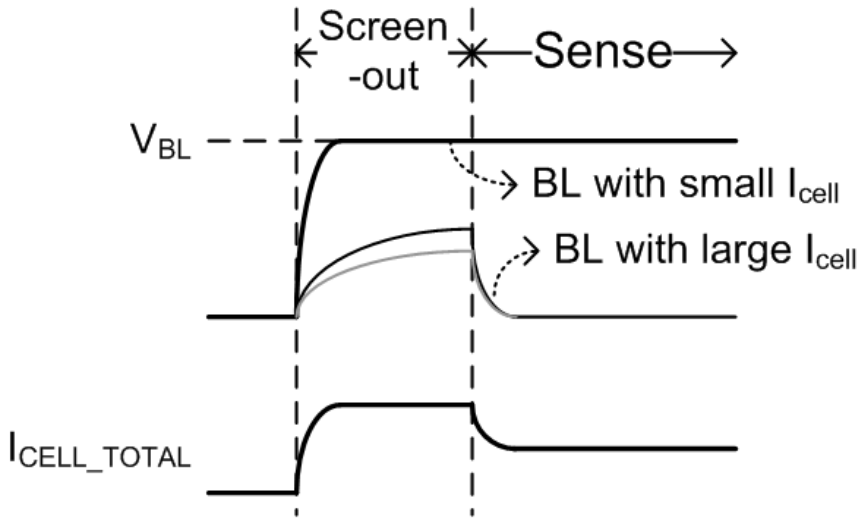
- Effect of remained cells during the sensing operation
  - Accuracy degradation
  - Large power consumption



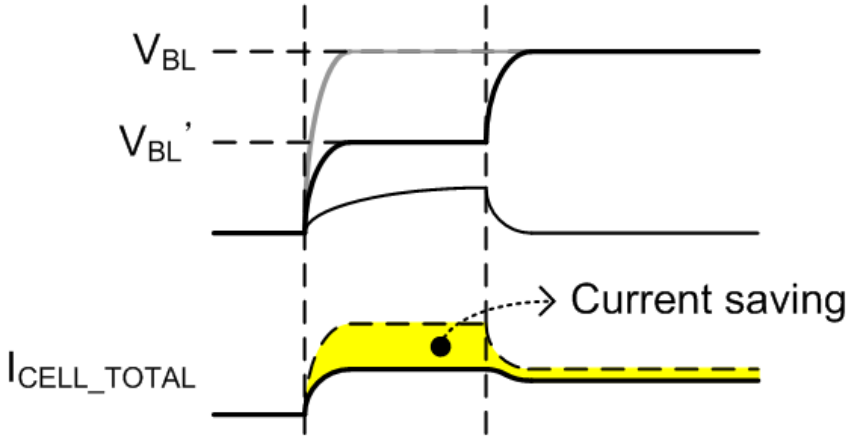
I_{cell}-controlled screen-out sensing

- Reduce I_{cell} during screen-out operation
  - Low V_{BL} (=V_{BL}') reduces I_{cell}.
- Recover I_{cell} during sensing operation
  - Recover V_{BL} level and sensing.





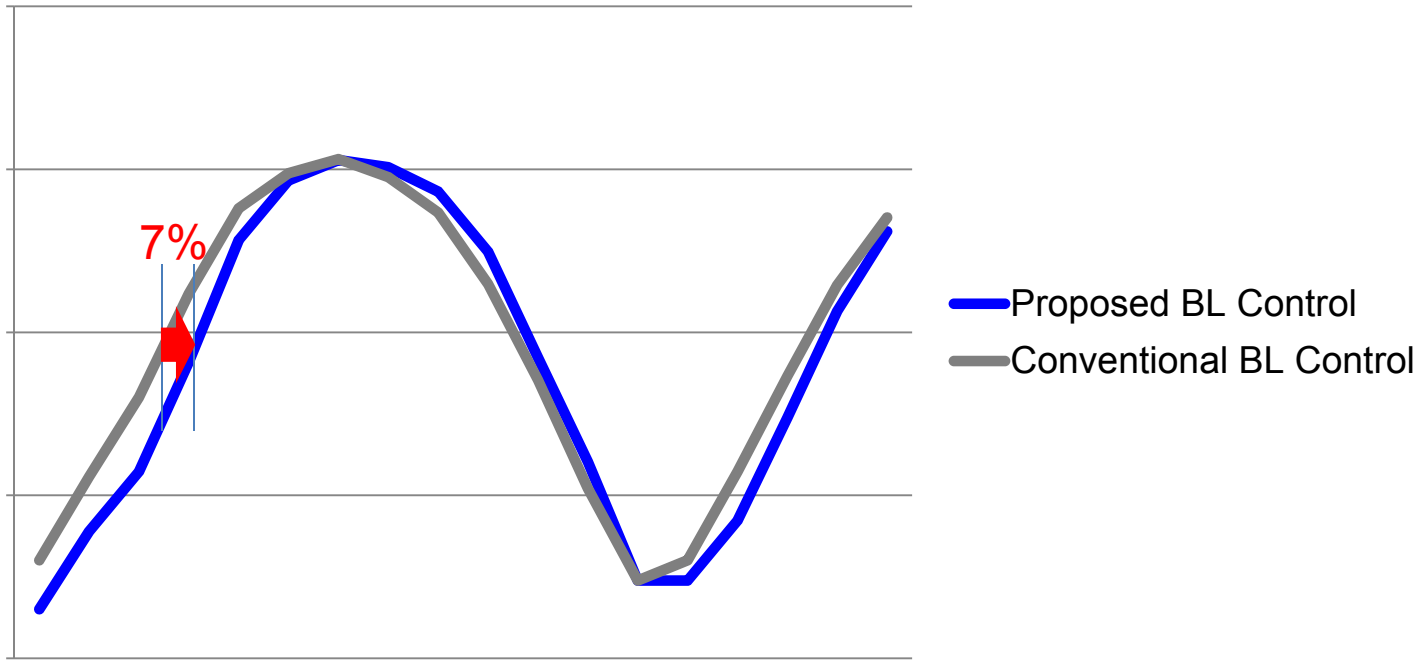
Conventional BL Control



Proposed BL Control

Measured Results

- Cell-current-controlled screen-out sensing scheme achieves 7% narrower cell distribution.

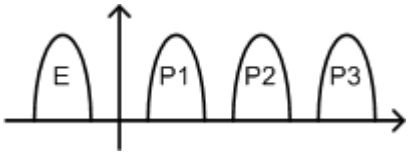


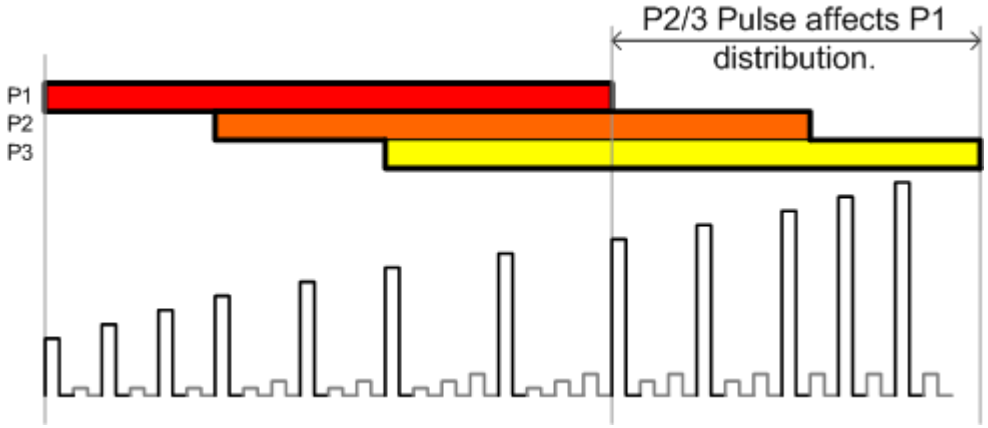
PGM Pulse for MLC Distribution

* Incremental Step Pulse Programming

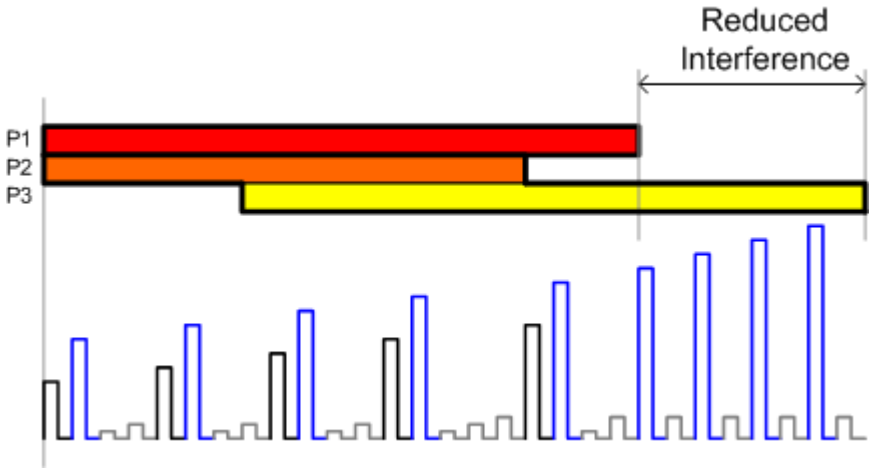
- ISPP* reduces the distribution of each cell Vth including P1, P2 and P3.

- Parallel PGM pulse scheme reduces P1 distribution caused by P2&3 PGM pulse.





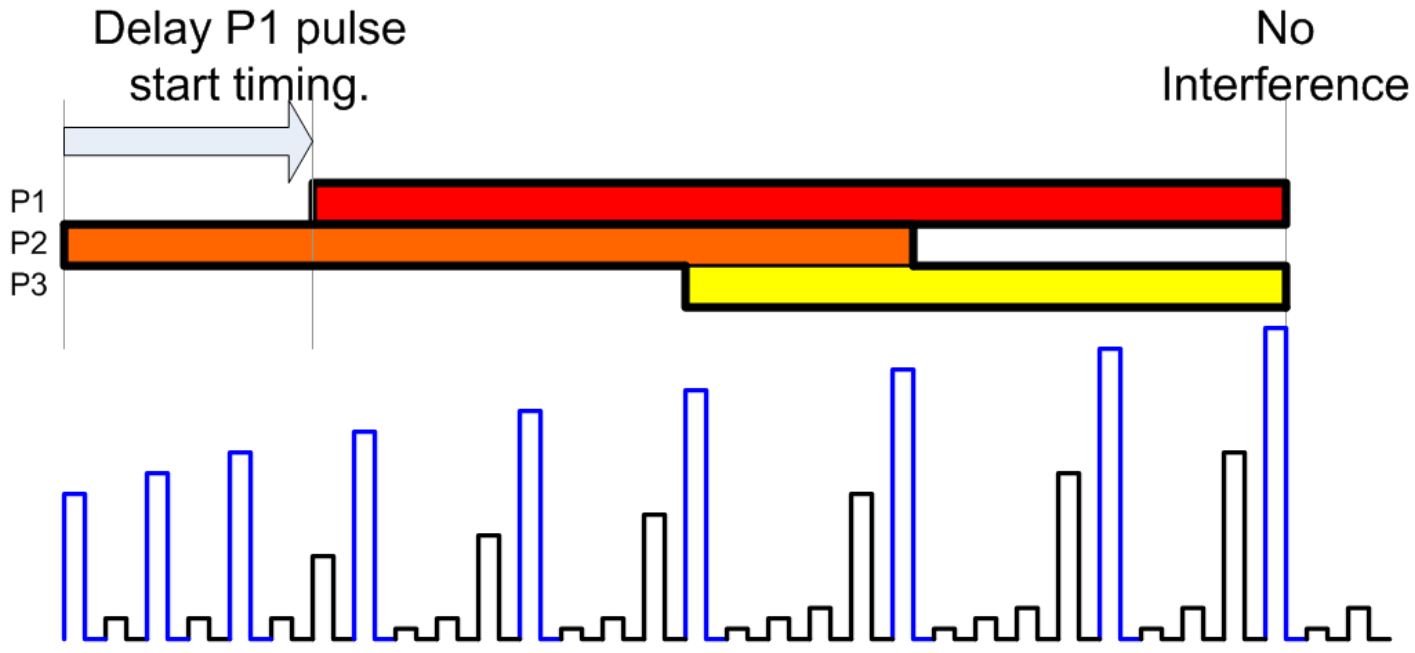
Conventional PGM Pulse (ISPP)



Parallel PGM Pulse

Delayed P1 PGM Pulse

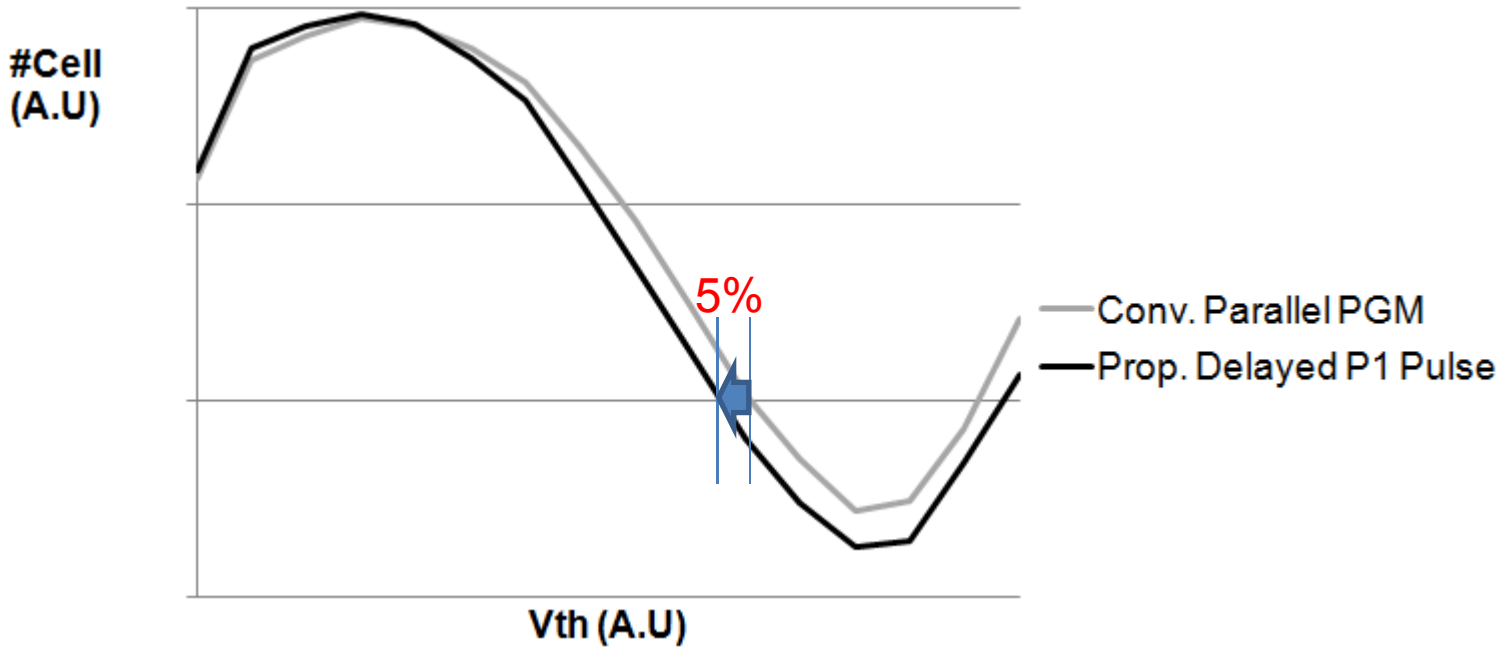
- Proposed delayed P1 PGM pulse prohibits wide P1 distribution caused by P2&3 PGM pulse.



Proposed Delayed P1 PGM Pulse

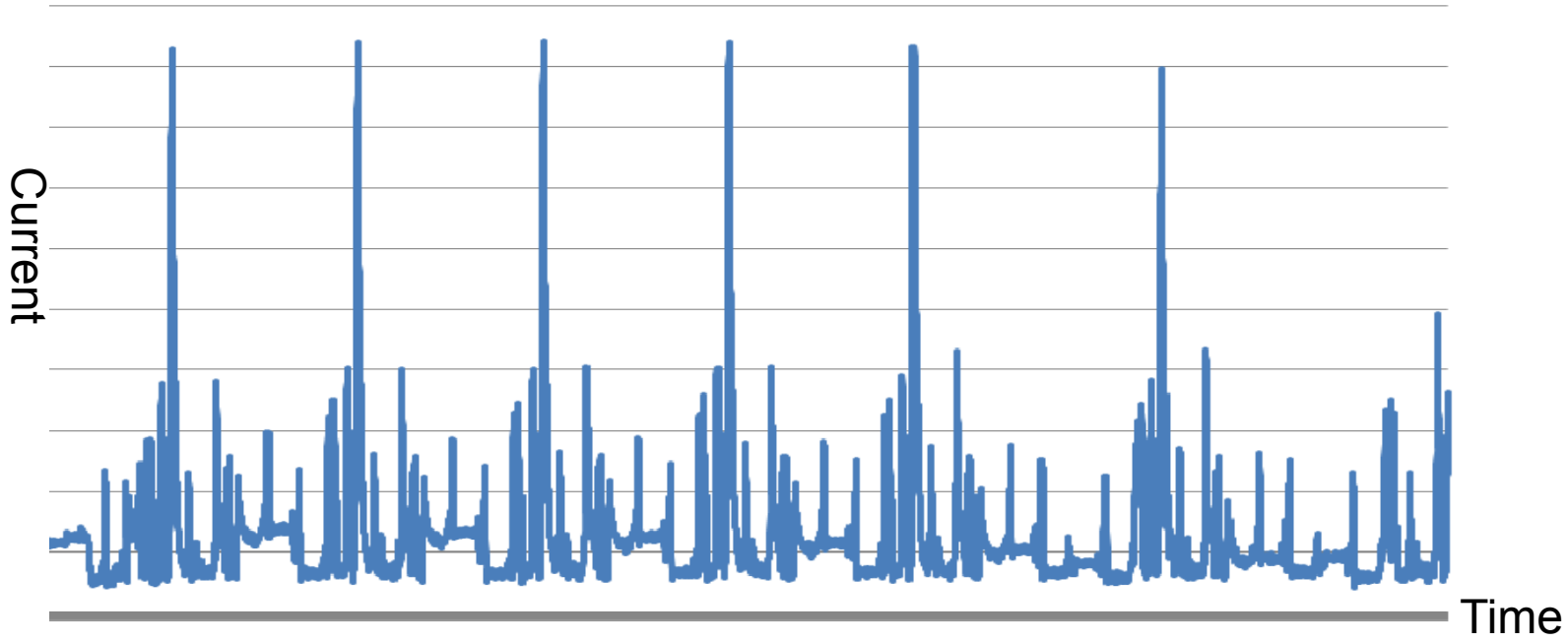
Distribution Comparison

- 5% narrower P1 distribution than conventional parallel PGM approach.



Peak Current Control

- Large peak current aggravate the supply voltage level stability.
  - Large decoupling capacitor for compensation
  - Large silicon area for large decoupling capacitor

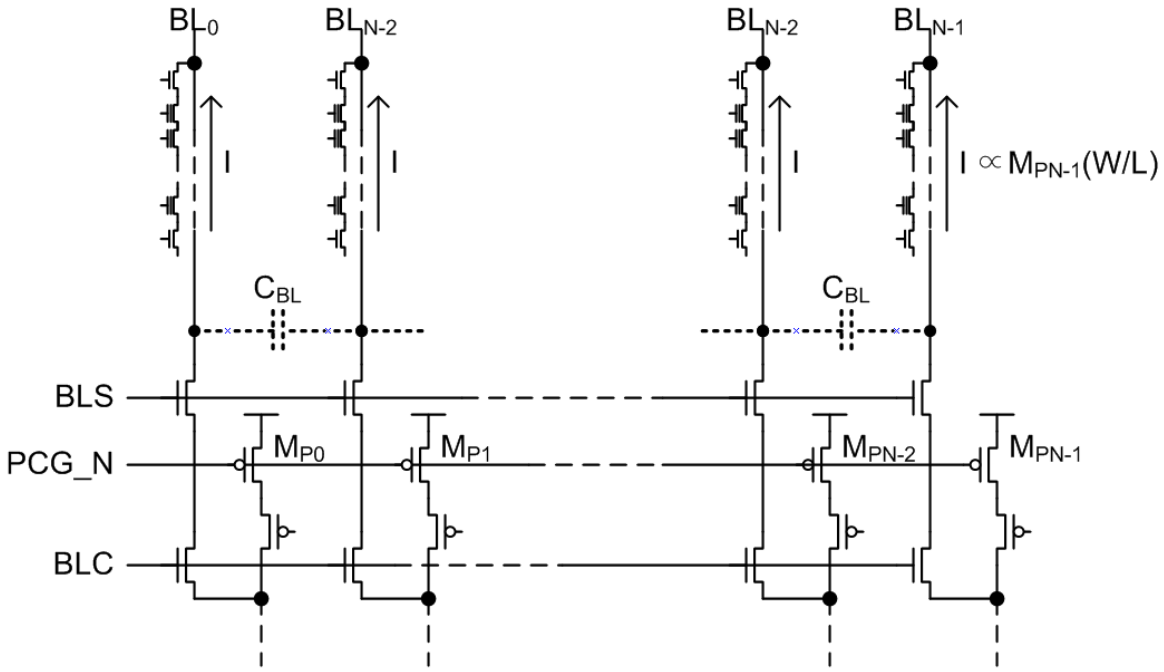


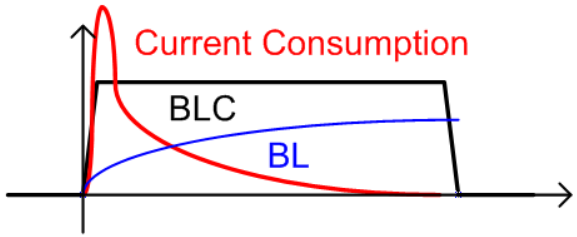
NAND flash current consumption waveform

Conventional BL Charging

- At the start of BL pre-charge
  - M_{Pi} flows its max amount of current.
  - Occurs current peak.

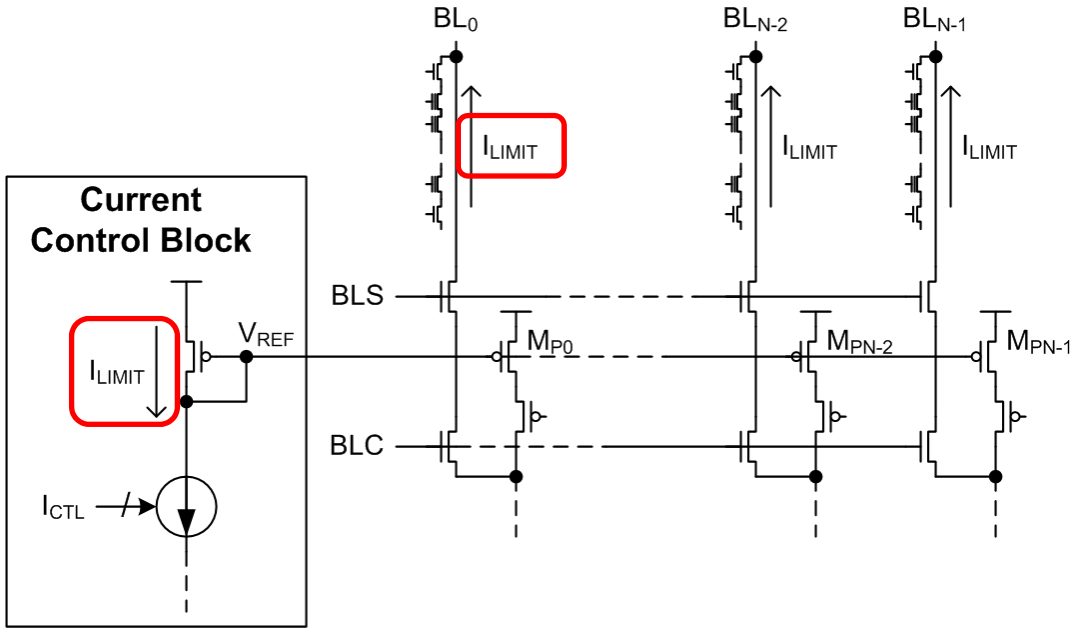
- During the BL pre-charge
  - Amount of current is scaled-down due to M_{Pi} level.

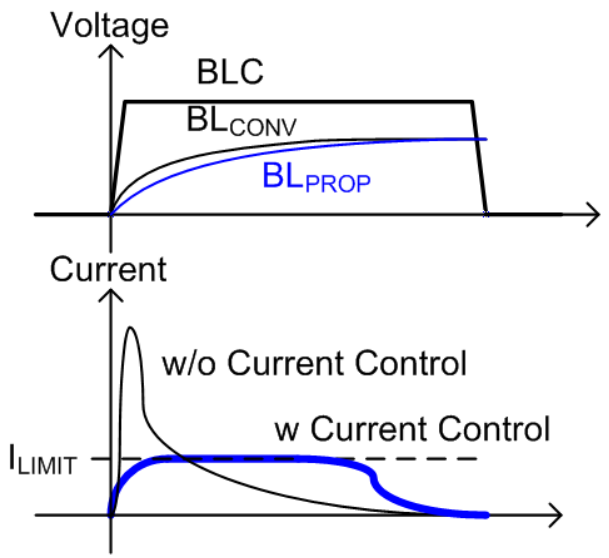




Proposed BL Charging

- Current control block limits max. BL pre-charge current.
  - Max value is configured by I_{CTL} binary value.
  - Peak current = # precharged BL * I_{LIMIT}

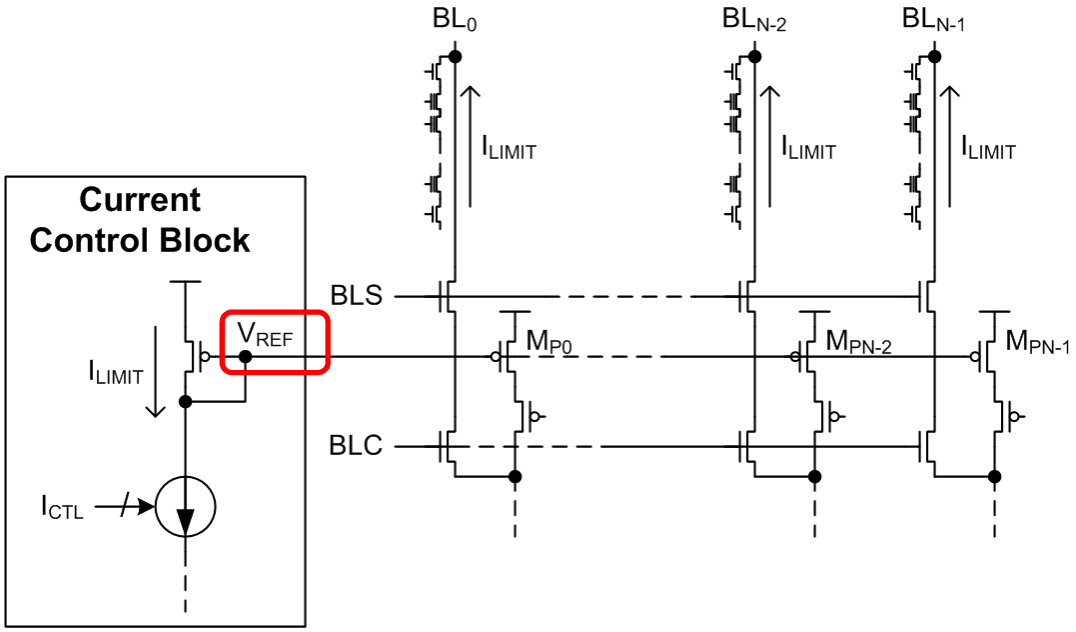


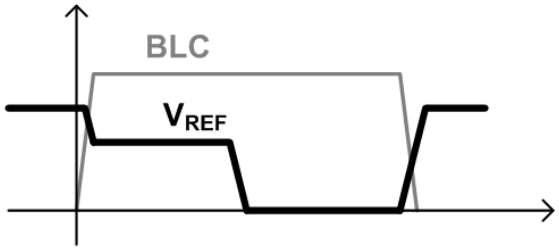


V_{REF} Control for Min. Timing Penalty

- Keeping V_{REF} = I_{LIMIT} causes longer BL precharge time.

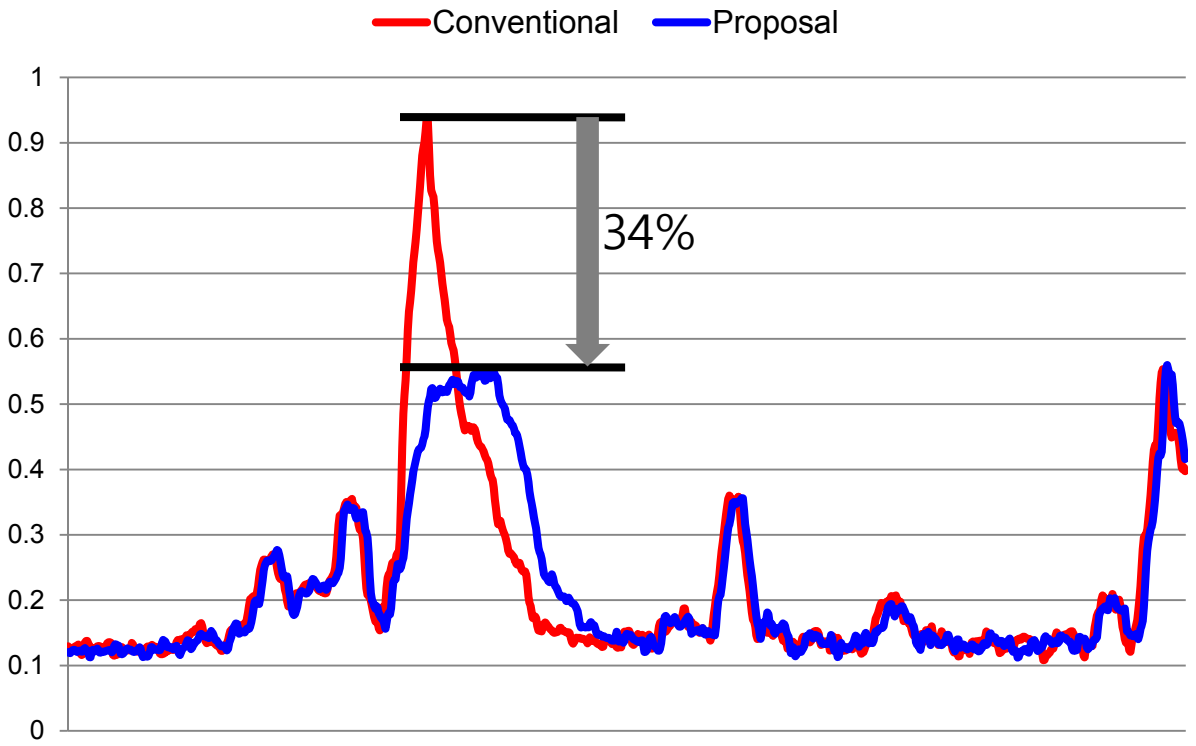
- V_{REF} = I_{LIMIT} during pre-defined time slot.
  - Time slot depends on the amount of I_{LIMIT}.
  - Time slot depends on Max. I_{peak} spec.





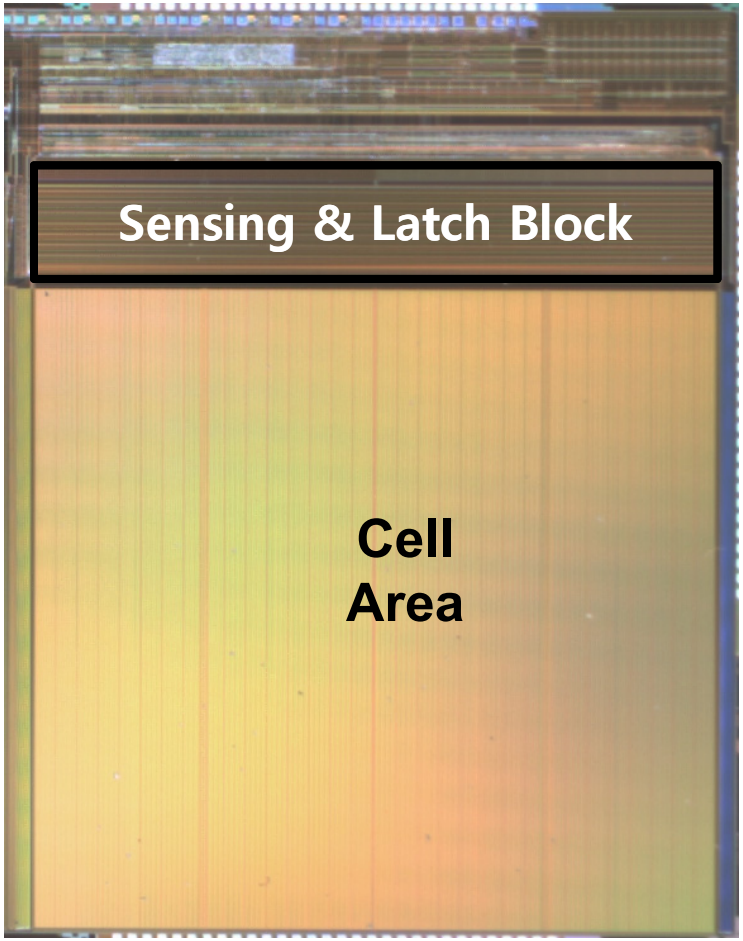
Peak Control: Measured Results

- 25~40% peak current reduction
  - 34% reduction with random data pattern

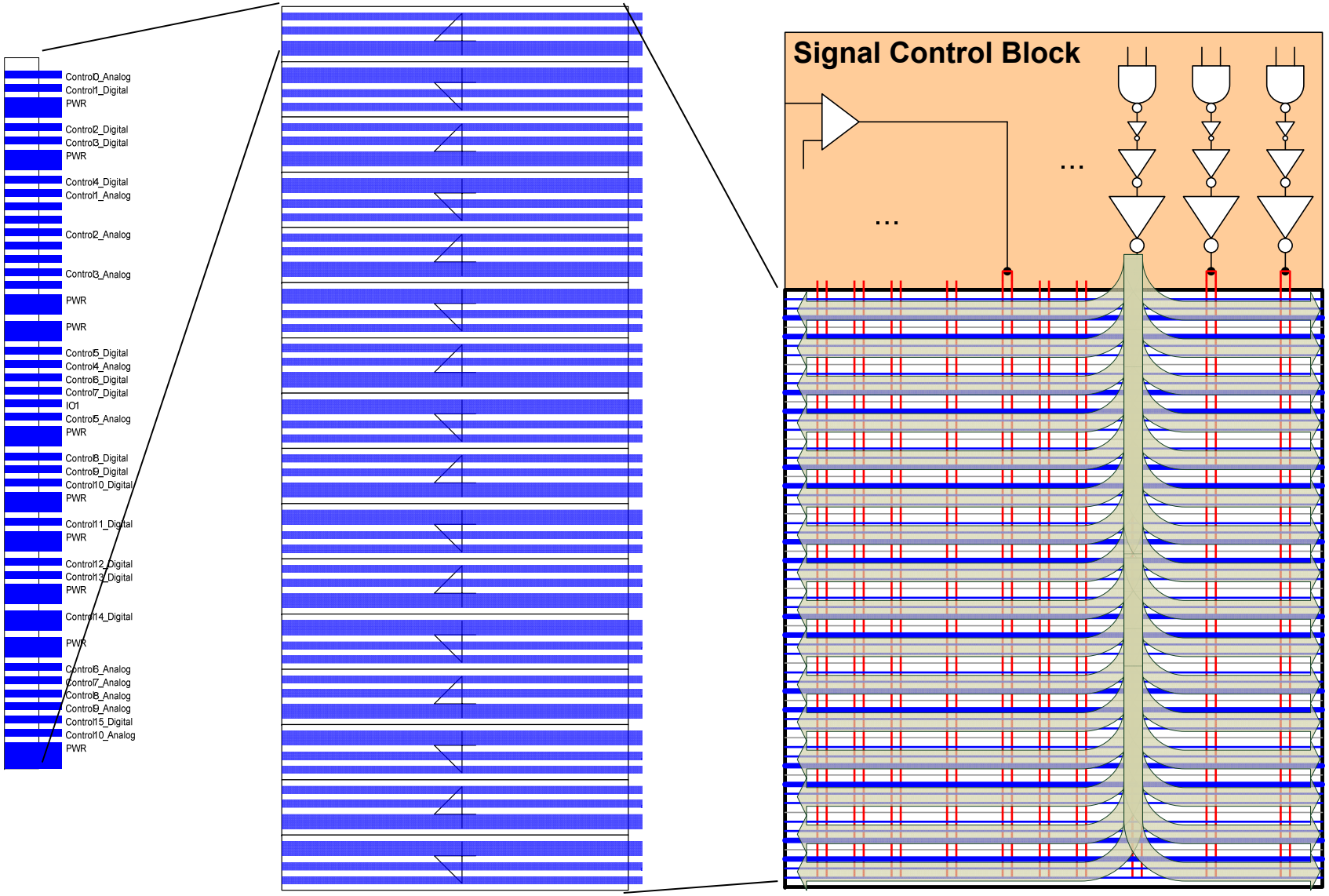


Load Balancing for Control Signals

- Many control signals for Sensing & Latch block
  - Digital Signals
    - Latch Control
    - Sensing Timing Control
  - Analog Signals
    - BL level
    - Sensing level
    - Current limit
  - High-Voltage Signals
    - NMOS pass-gate

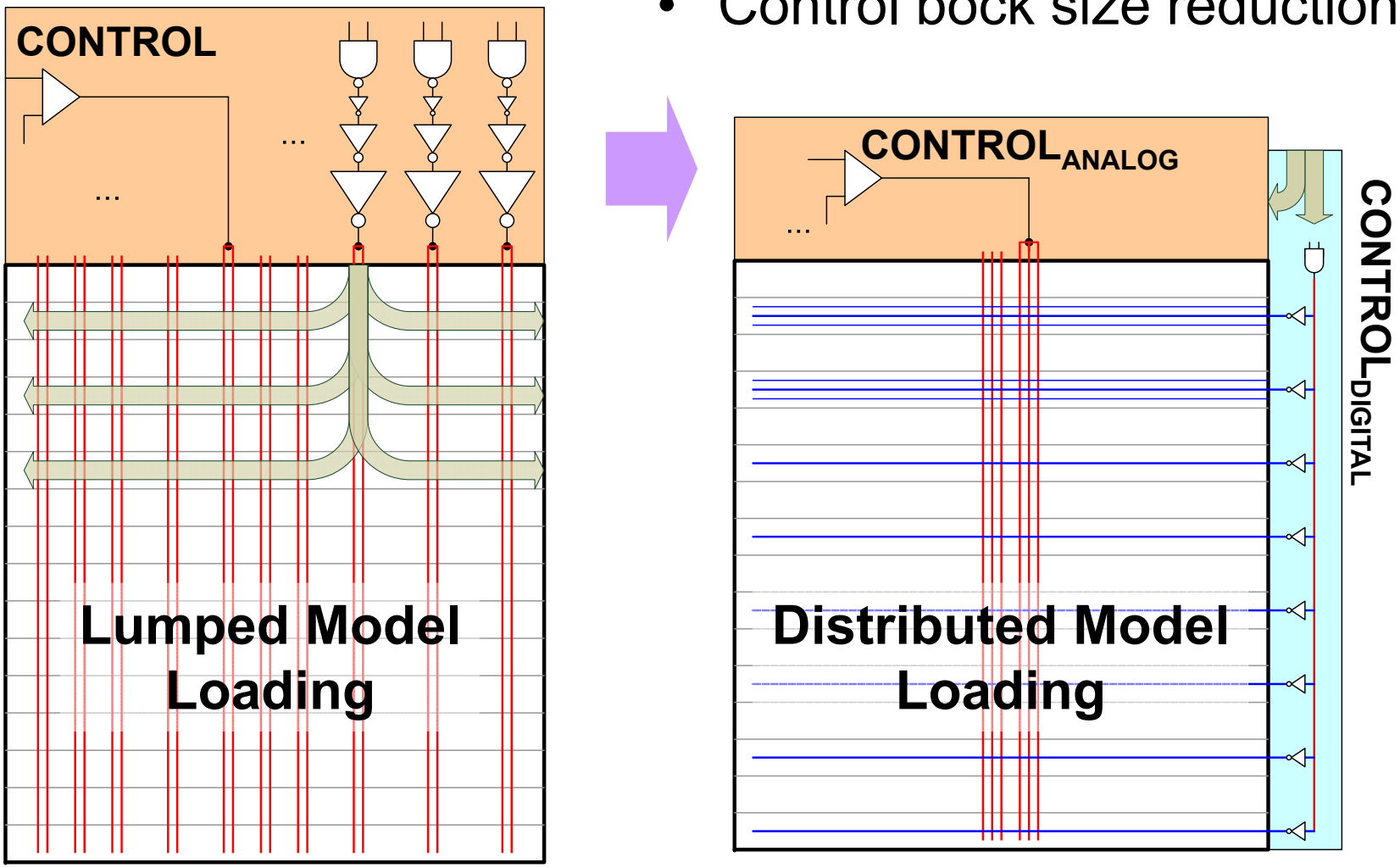


Control Signals P&R

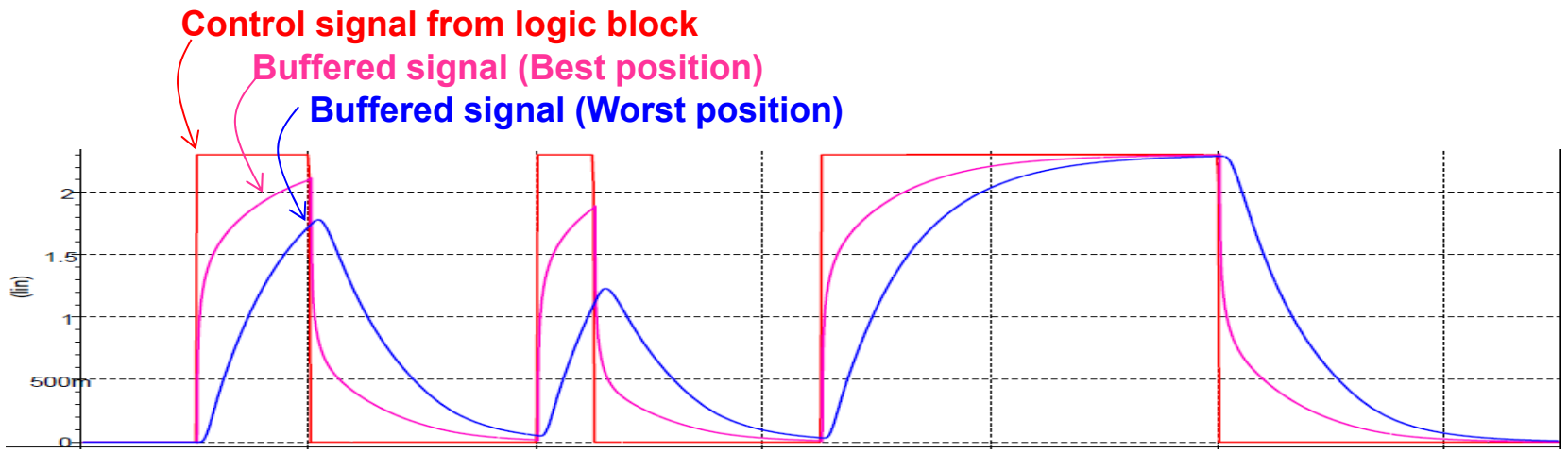


Load-balanced control signal P&R

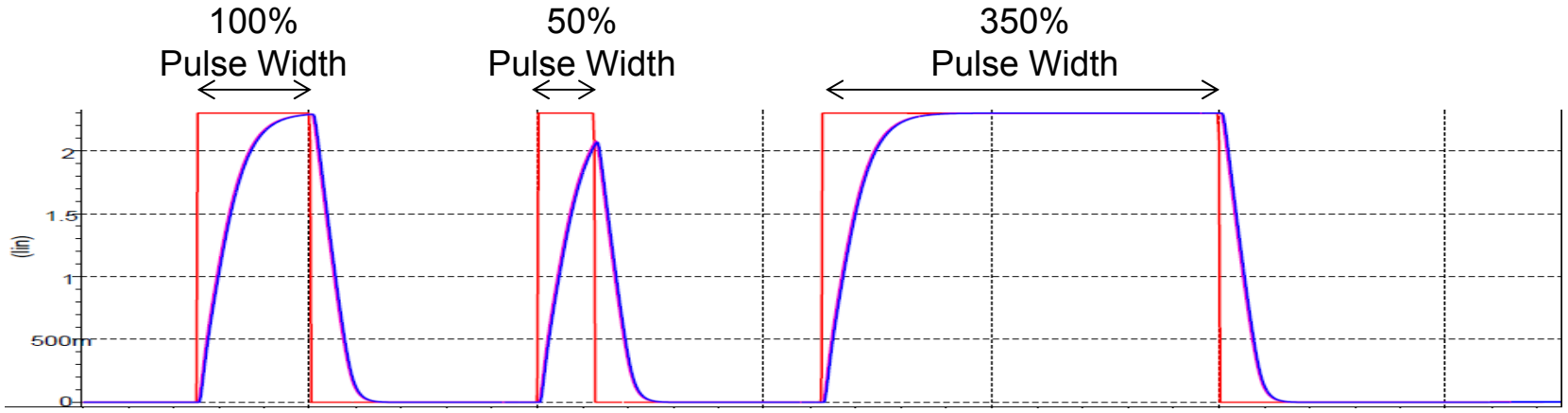
- Relieve loading.
- Control bock size reduction



Simulation Results



Conventional Control Block

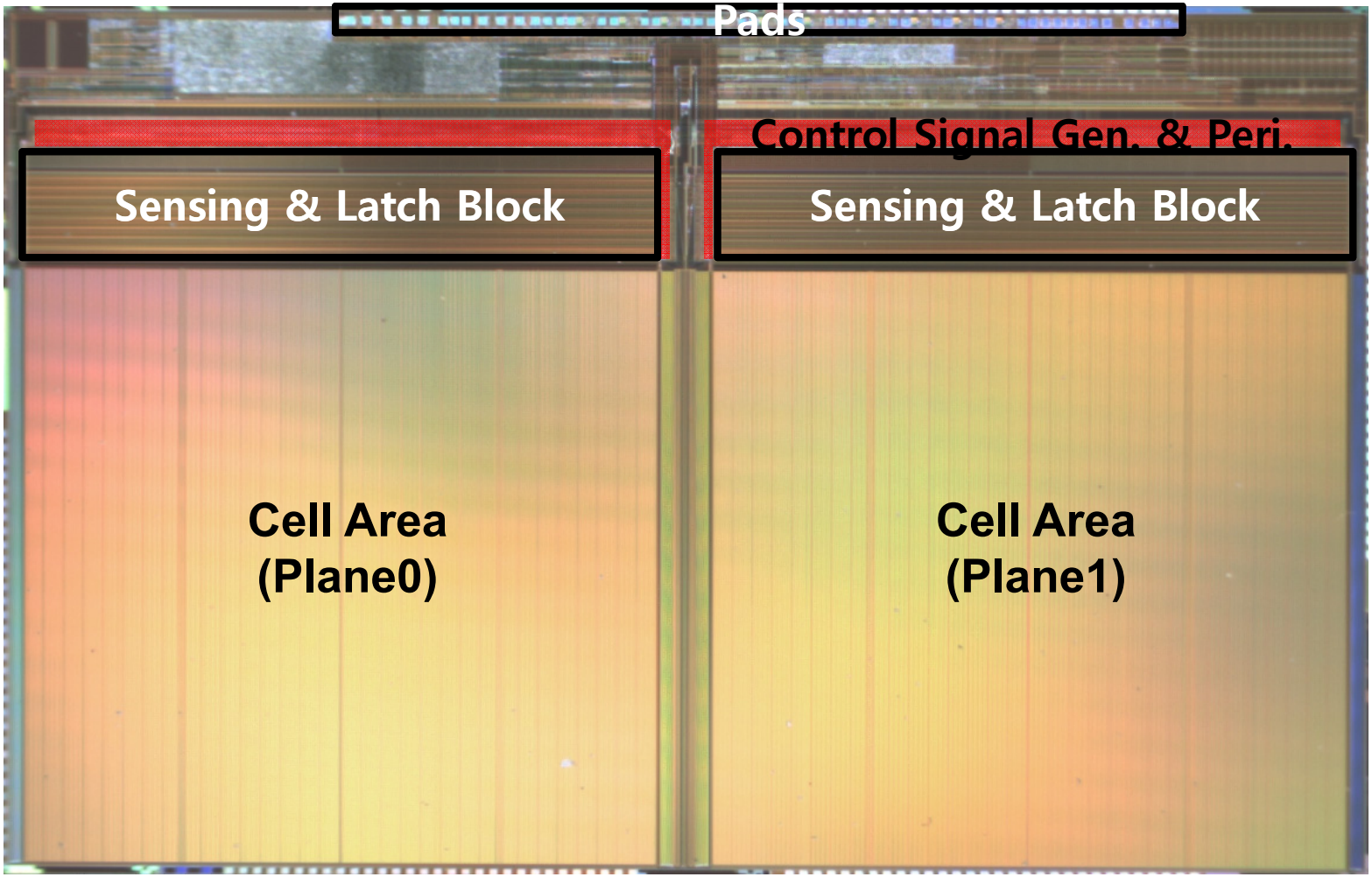


Proposed Control Block

Chip Key Features

Technology	16nm 3-Metal CMOS
Density	64Gb (2bit/cell)
Organization	16K bytes x 256 pages x 1K blocks x 2 planes
Block Size	4MB
Power Supply	2.7V ~ 3.6V
Data Transfer Rate	400Mb/s
Program Throughput	25MB/s (Typical)

Die Micrograph



Summary

- 1st & smallest 64Gb NAND flash memory with 16nm floating-gate cell structure
- Cell-current-controlled screen-out sensing scheme both reduces power consumption and enhances sensing accuracy.
- Delayed P1 PGM pulse scheme narrows P1 distribution.
- Peak-current controller reduces 25~40% of peak current consumption.
- Load-balanced control signal P&R enhances control signal integrity.

66.3KIOPS-Random-Read 690MB/s-Sequential-Read Universal Flash Storage Device Controller with Unified Memory Extension

Konosuke Watanabe[1], Kenichiro Yoshii[1], Nobuhiro Kondo[1], Kenichi Maeda[1], Toshio Fujisawa[1], Junji Wadatsumi[2], Daisuke Miyashita[2], Shouhei Kousai[2], Yasuo Unekawa[2], Shinsuke Fujii[2], Takuma Aoyama[2], Takayuki Tamura[1], Atsushi Kunimatsu[1], Yukihito Oowaki[1]

Toshiba Corp., Semiconductor & Storage Products Company

{[1]Yokohama, [2]Kawasaki} - Japan

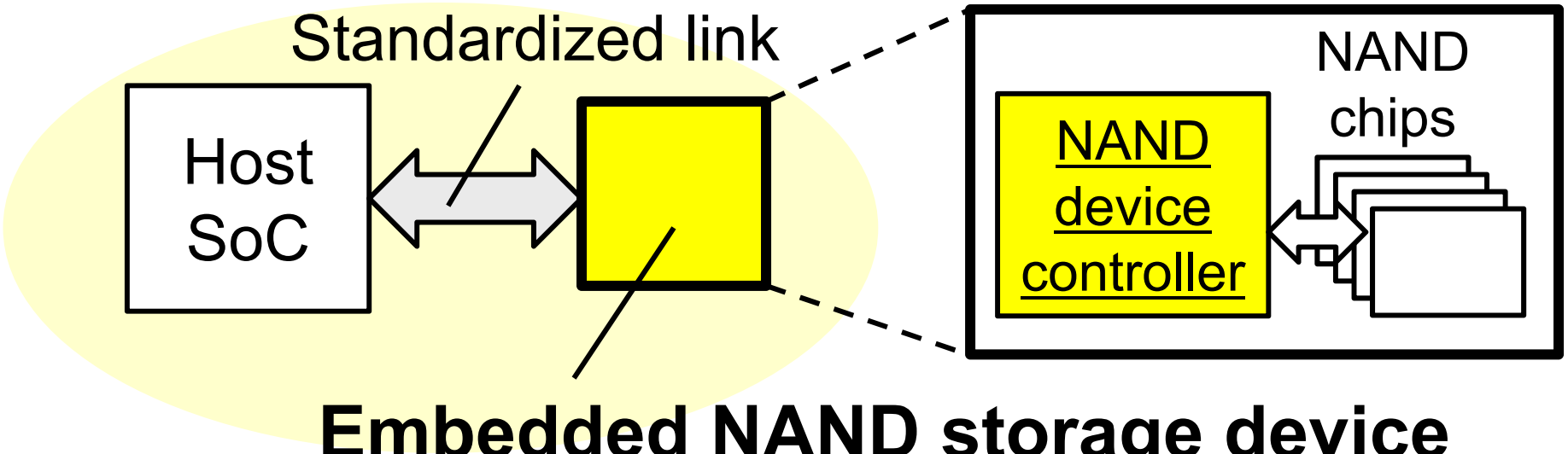
TOSHIBA
Leading Innovation

How did our embedded NAND storage device get SSD-like higher read performance?

Outline

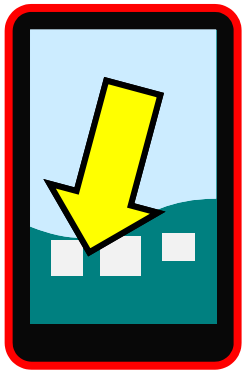
- Background and approaches
- Unified Memory Architecture
- Random Read Command Processor (RRCP)
- Synchronized-injection CDR
- Results

Embedded NAND Storage Device



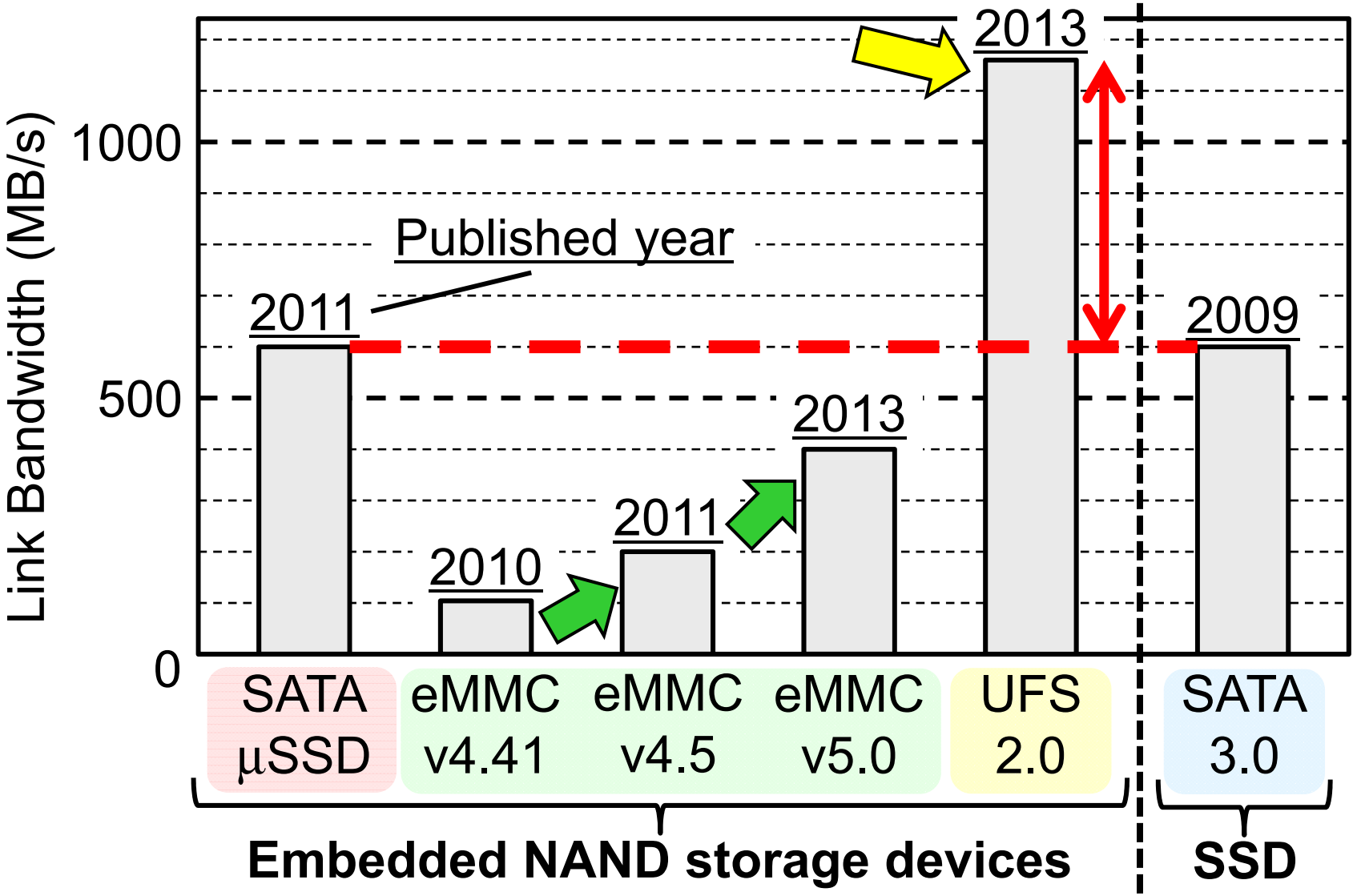
Embedded NAND storage device

- NAND storage designed for embedded system
- Single BGA package chip
  - Connects to host SoC with standardized link
  - Contains controller and NAND chips
- **Small and low power**
  - 10 - 20 mm / each side
  - 1 - 2 W (active)



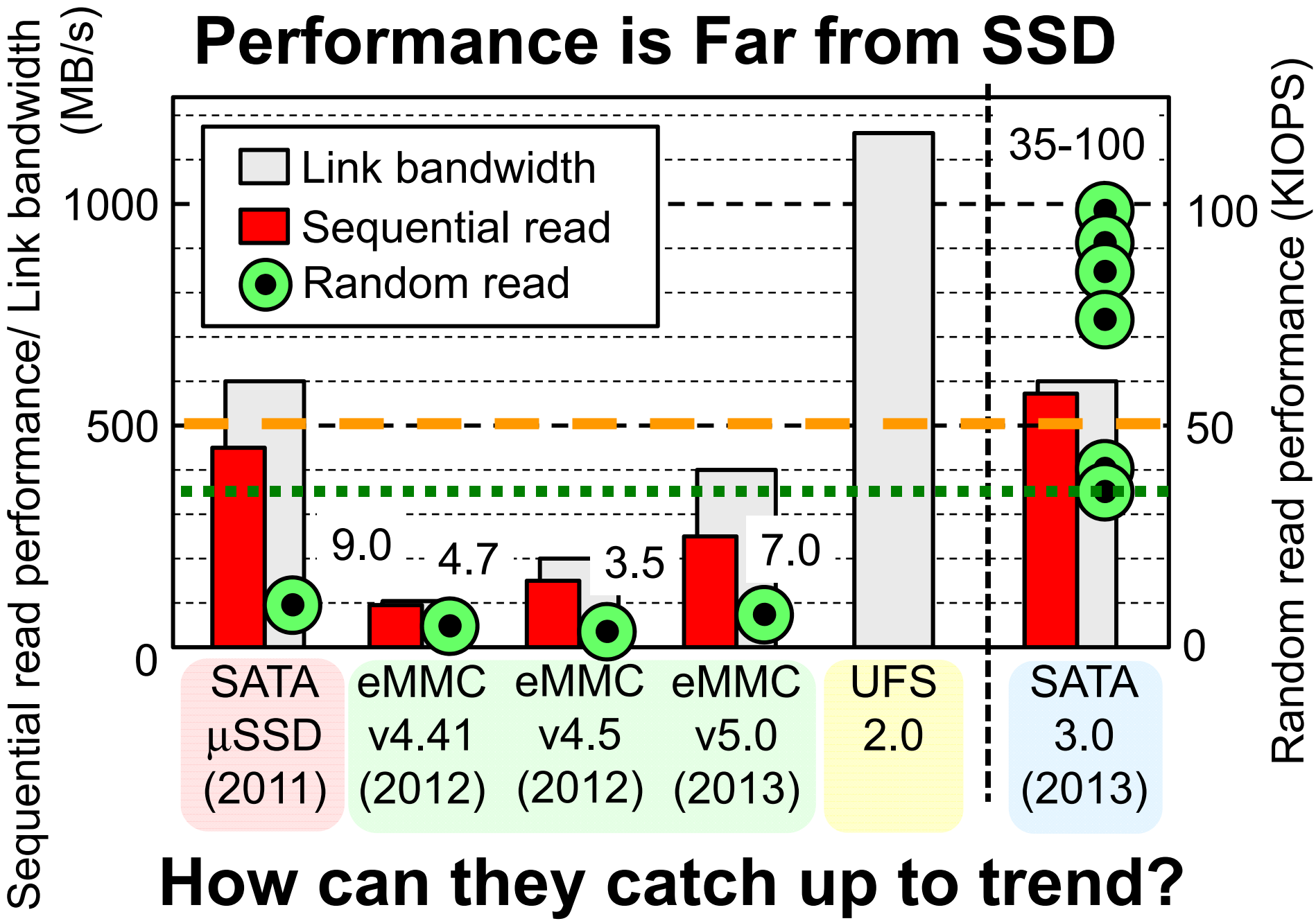
cf. SSD
- 2.5” form factor
- Rich power source

Link Bandwidths are Comparable to SSD



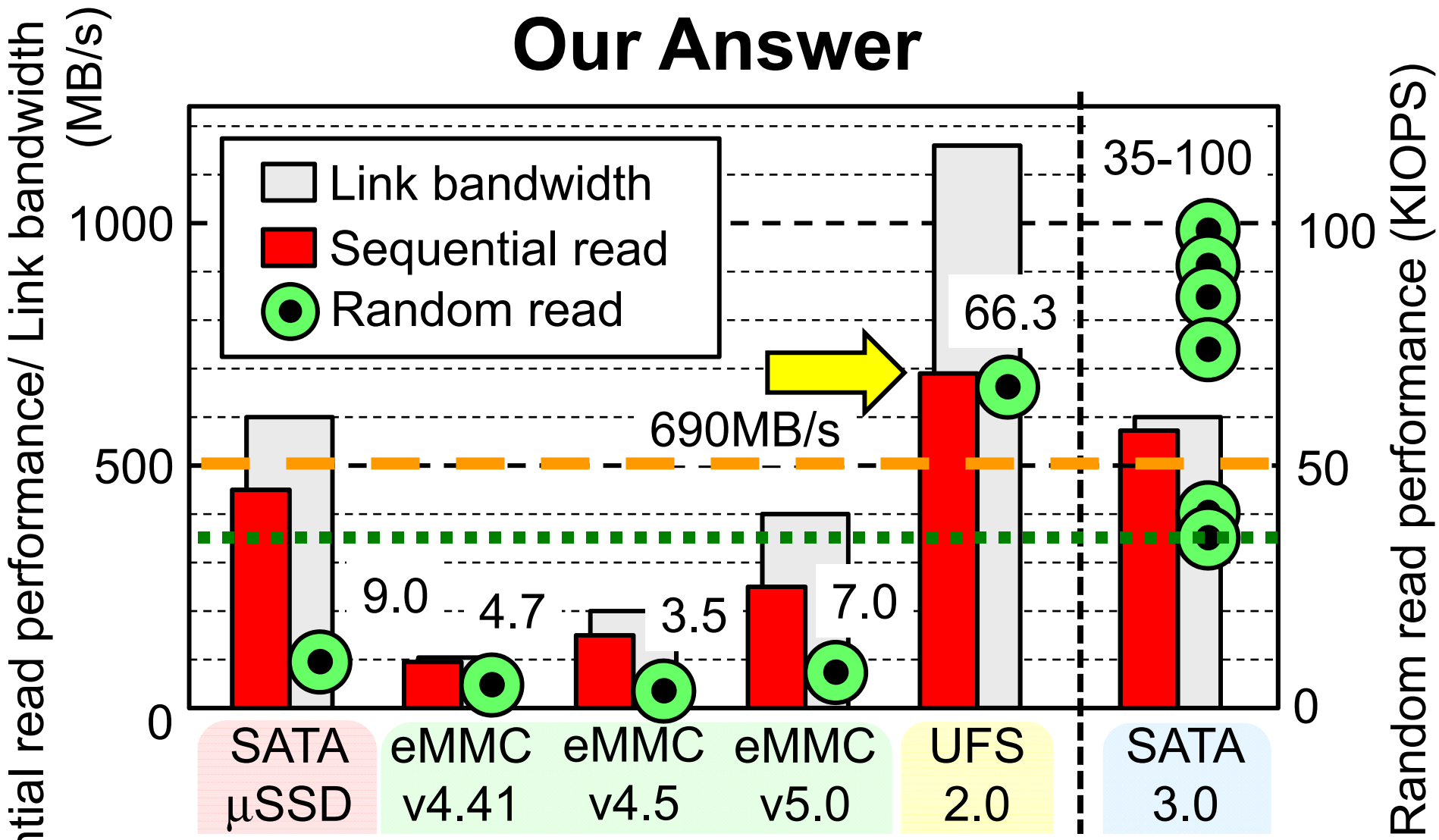
Standards are performance-oriented

Performance is Far from SSD



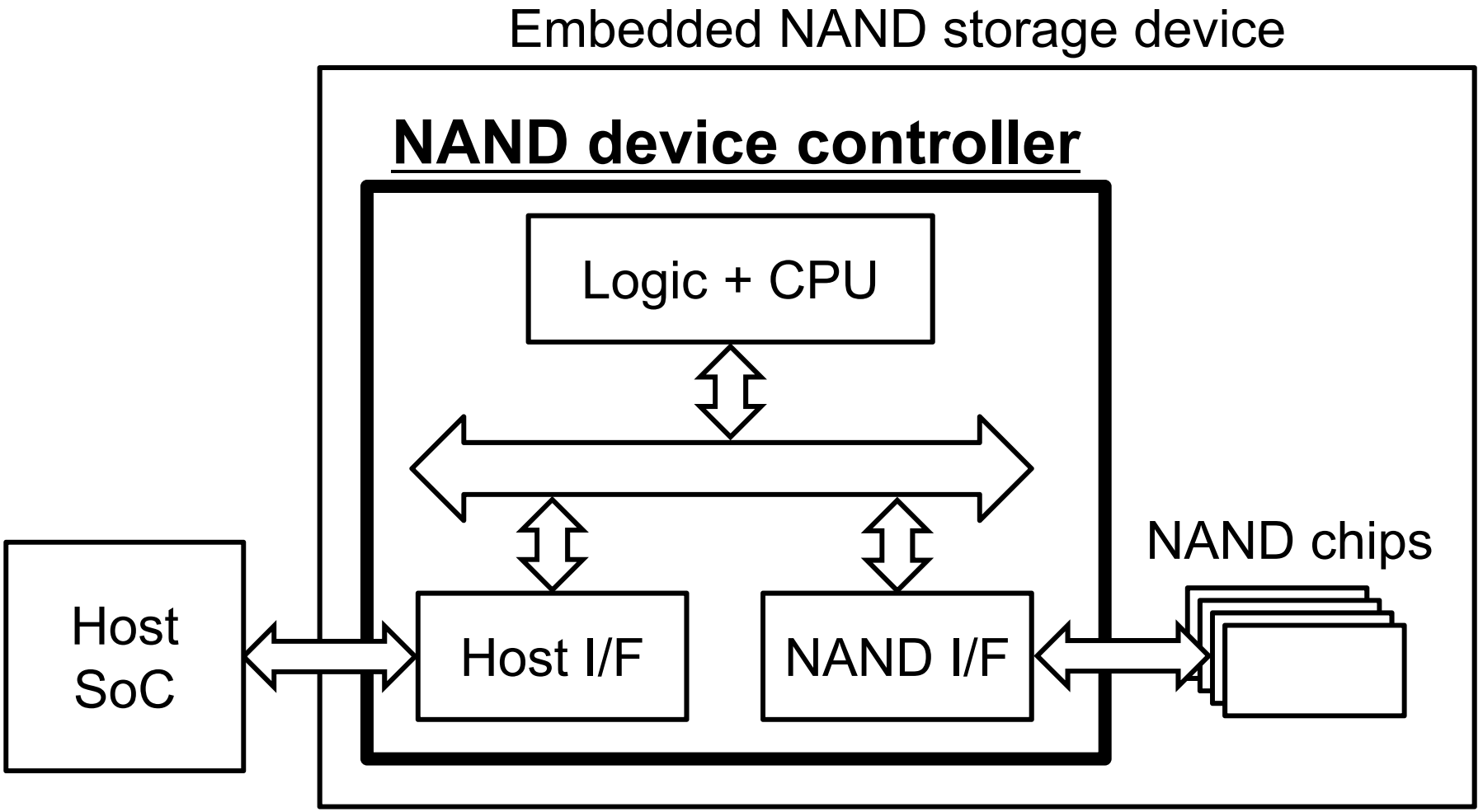
How can they catch up to trend?

Our Answer

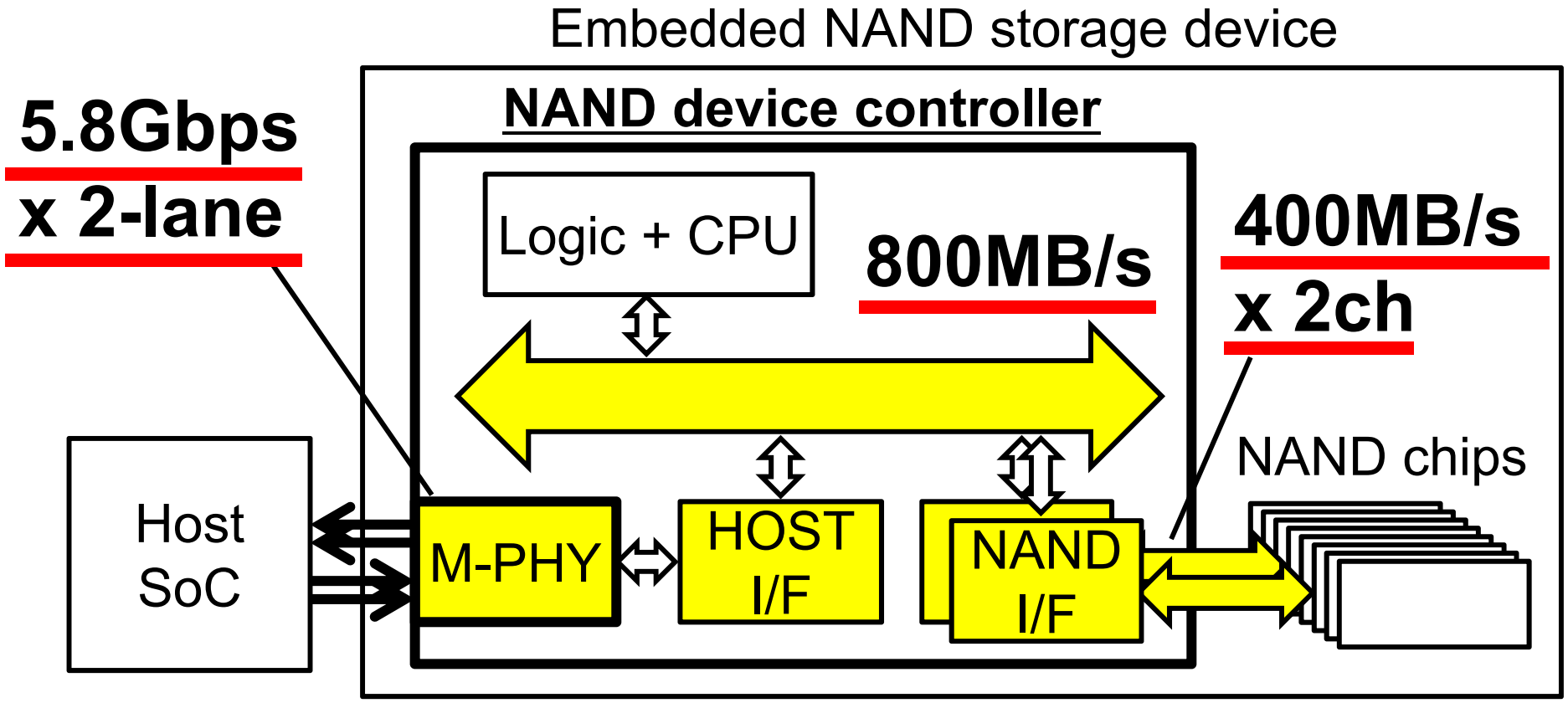


SSD-like higher read performance UFS 2.0 device

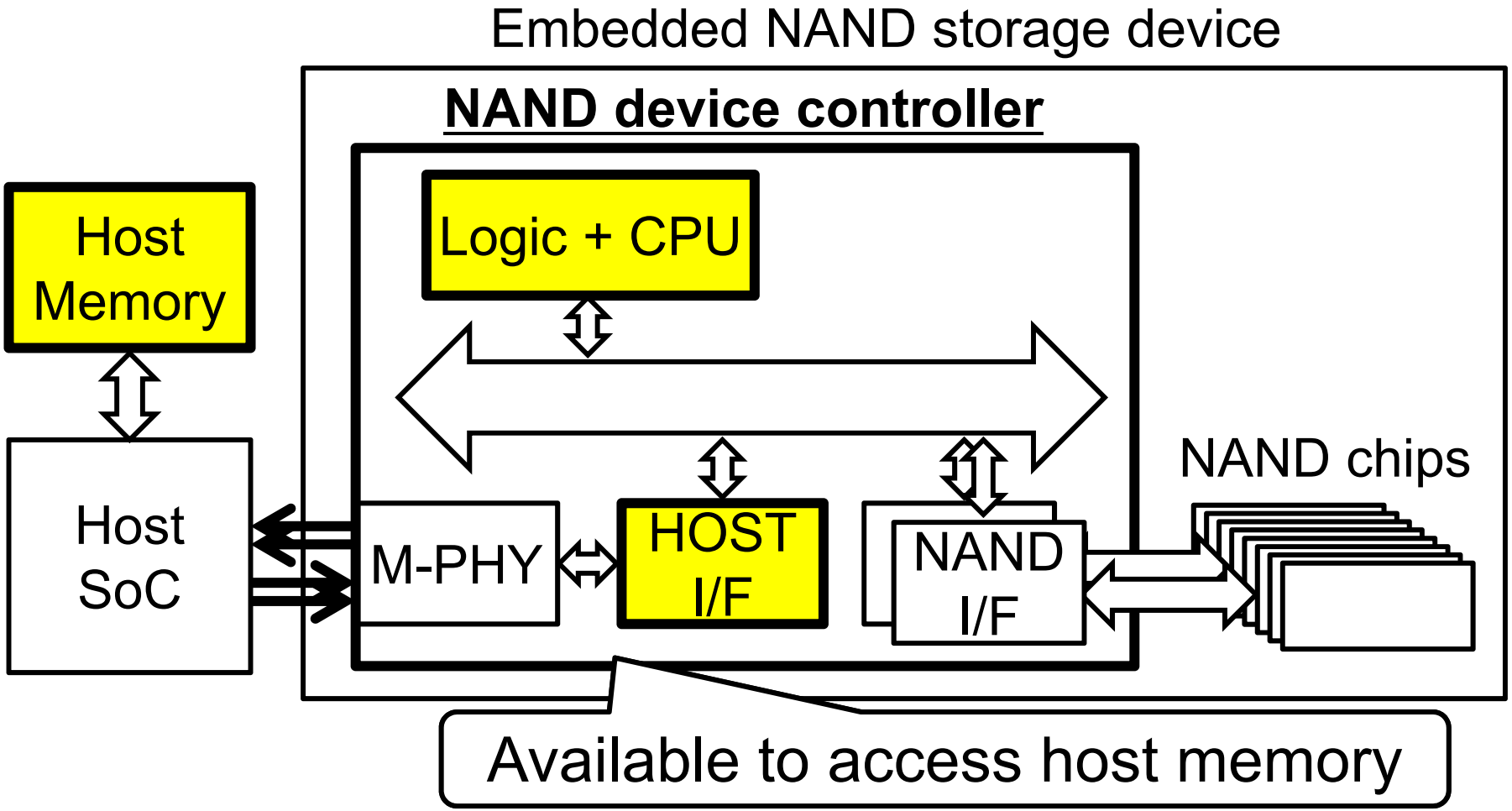
Plain NAND Device Controller of Embedded NAND Storage Device



To Get Sequential Read Performance
Strengthened Data Paths

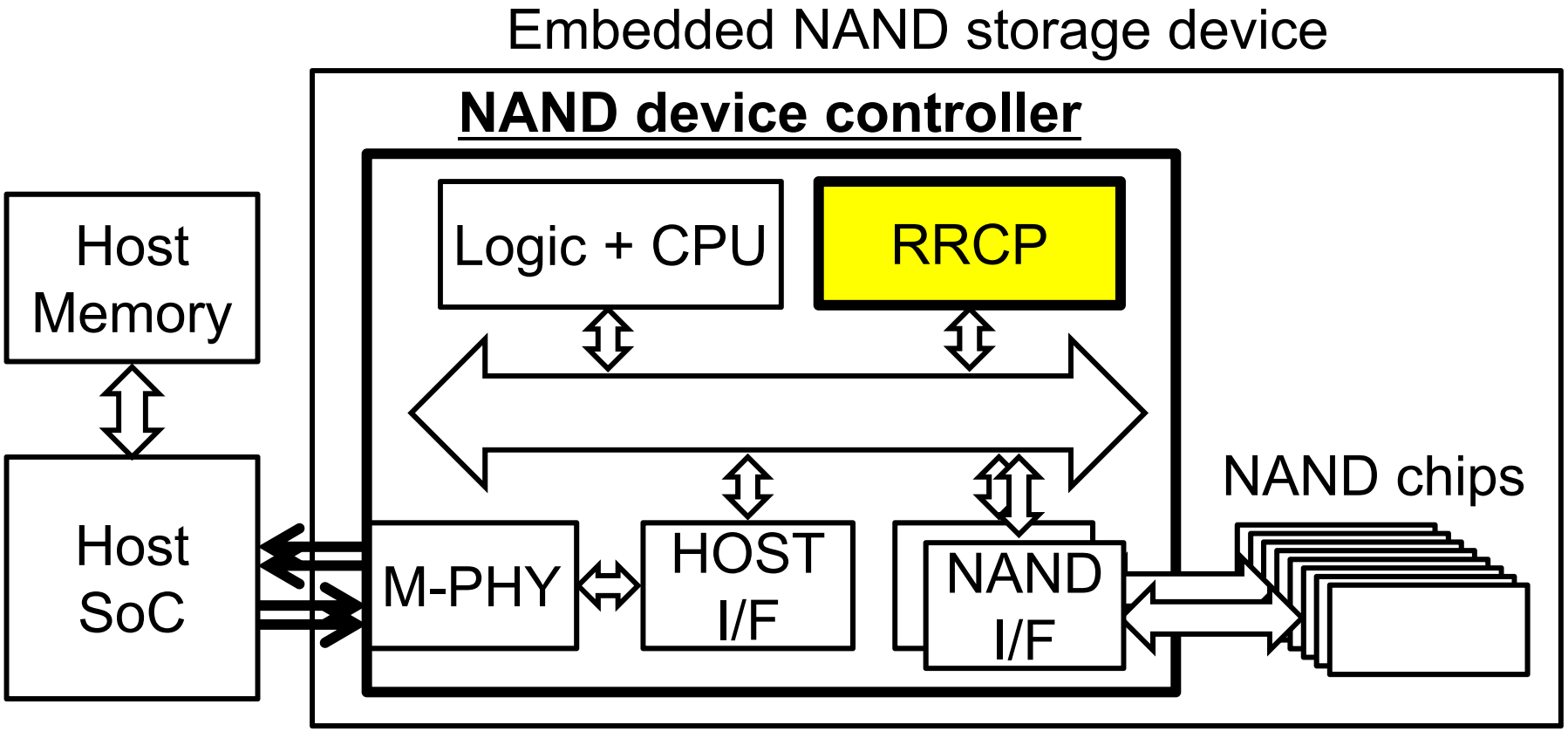


To Get Random Read Performance Introduced Unified Memory Architecture



Read latency became smaller

To Get More Random Read Performance Added Random Read Command Processor



Increased NAND parallelism

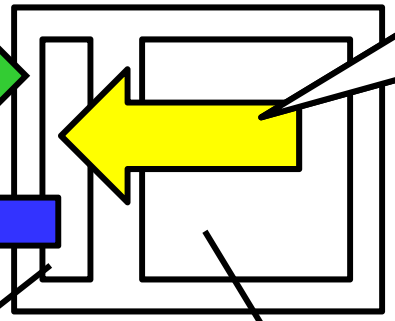
Outline

- Background and approaches
- Unified Memory Architecture
- Random Read Command Processor (RRCP)
- Synchronized-injection CDR
- Results

Page Read Latency of NAND Chip

1. Read command with address

< 10 cycle
= 0.05µsec (200MHz)



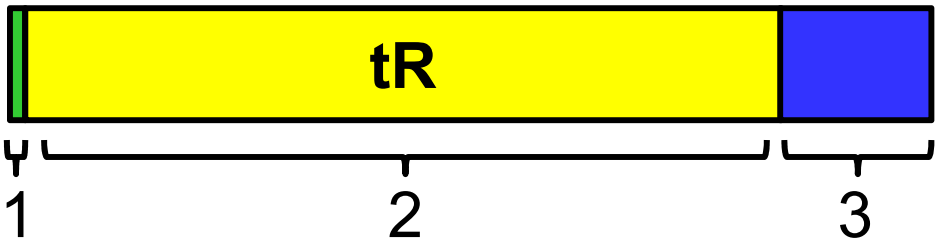
2. Transfers page data from NAND cell array to buffer

tR:
Several 10s µsec

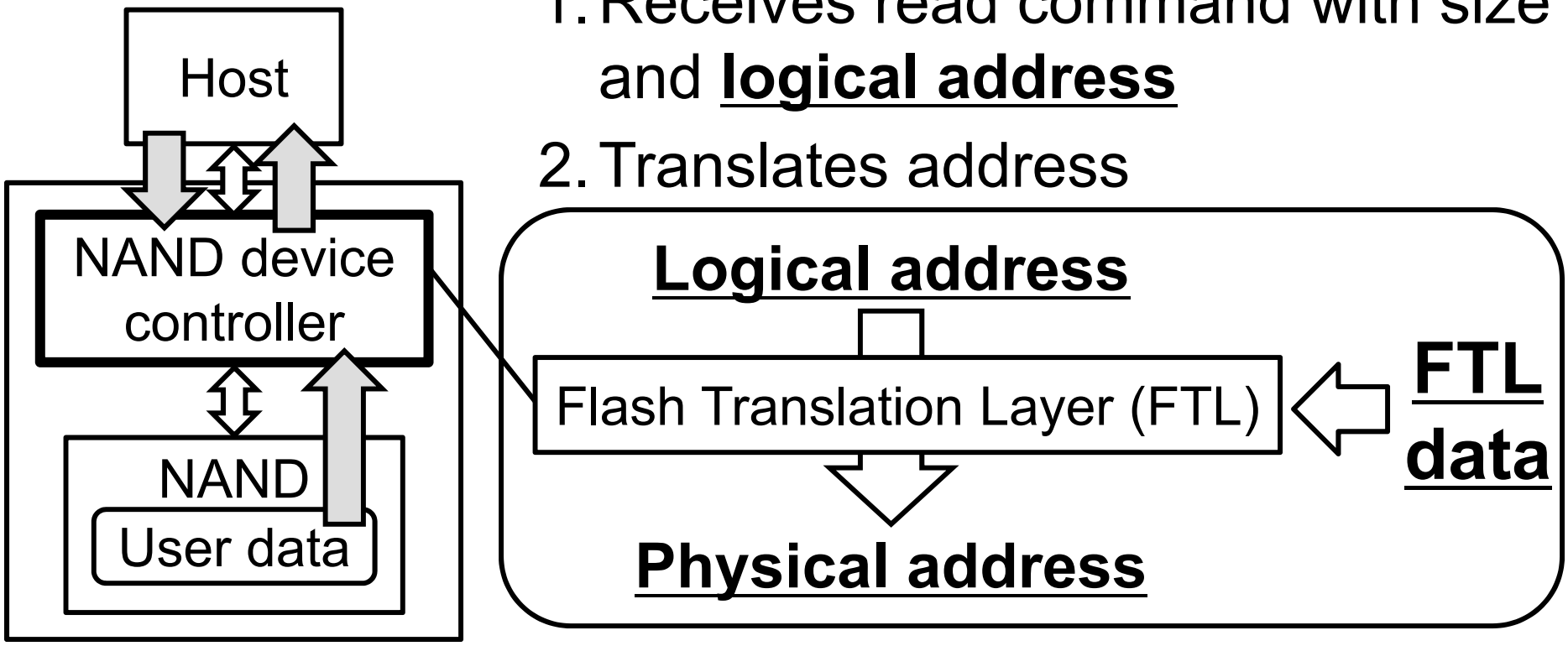
3. Data output

4096 + extra Byte (1 page = 4KB)
= 10µsec (400MB/s)

Larger than several 10s µsec

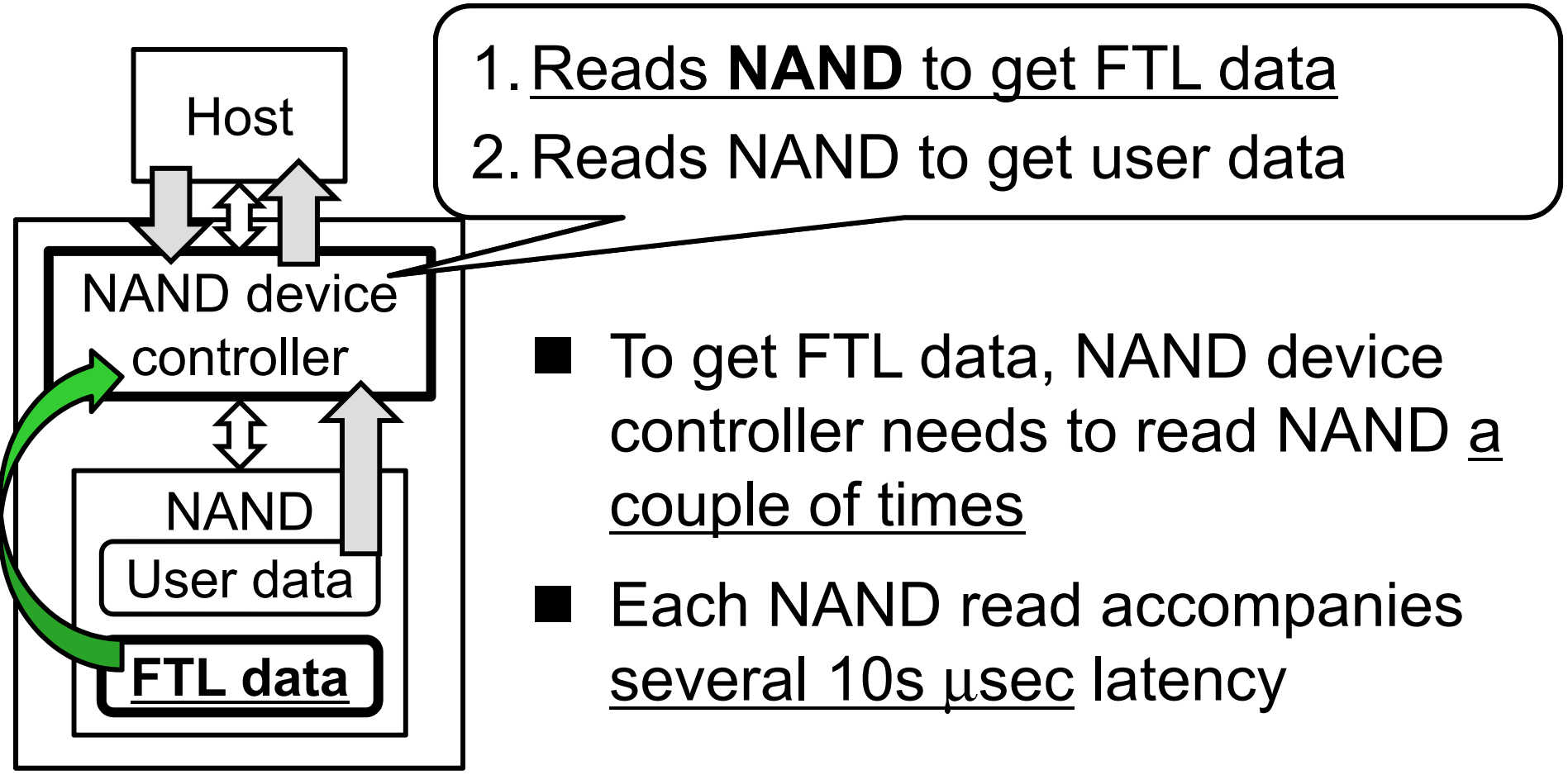


Flow of Random Read



FTL data affects to random read latency

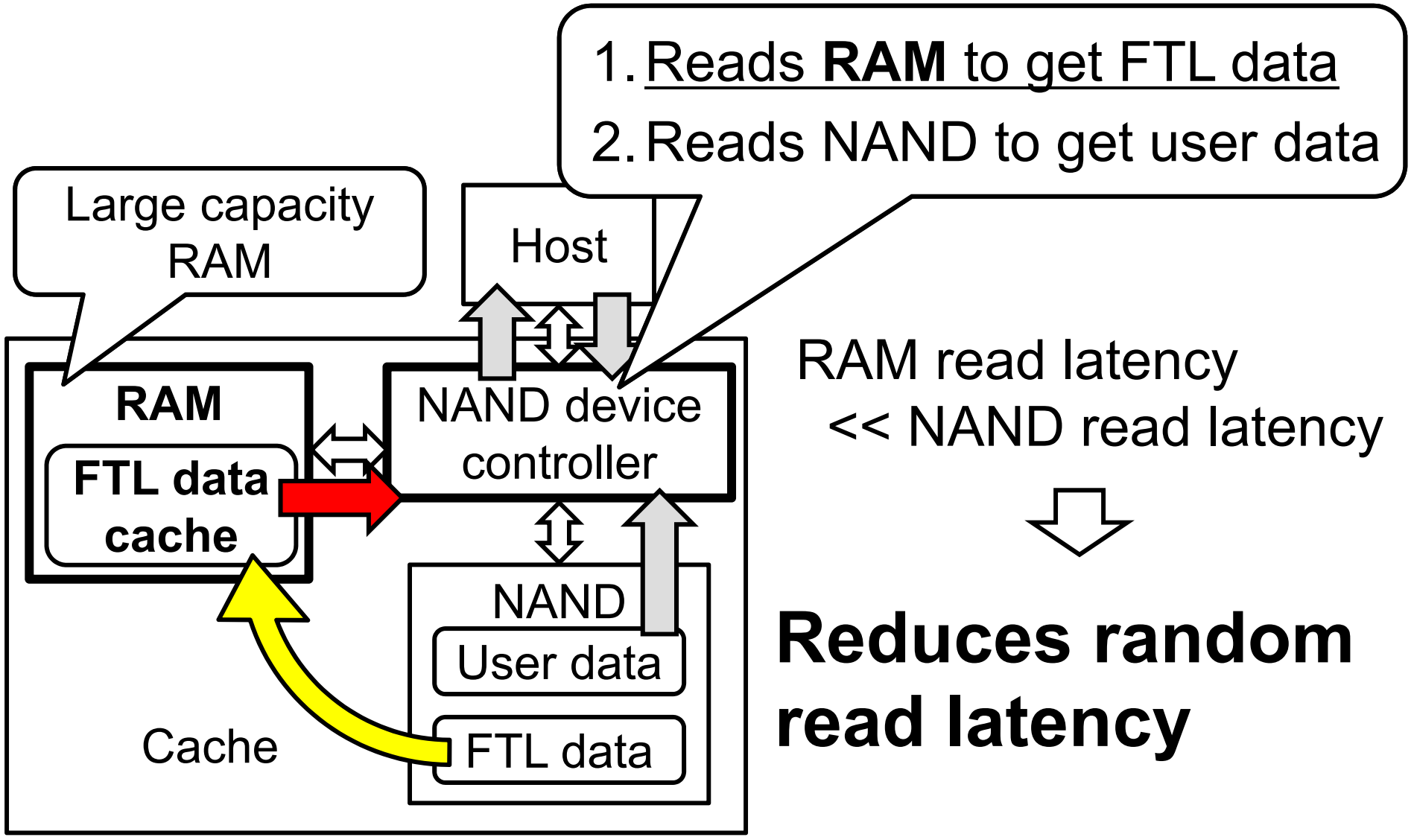
FTL Data is Also Stored in NAND Chips



- To get FTL data, NAND device controller needs to read NAND a couple of times
- Each NAND read accompanies several 10s µsec latency

FTL data read would increase random read latency more than 100µsec

On-chip RAM Solution



RAM read latency
<< NAND read latency

Reduces random read latency

On-chip RAM Solution is Unsuitable for Embedded NAND Storage Device

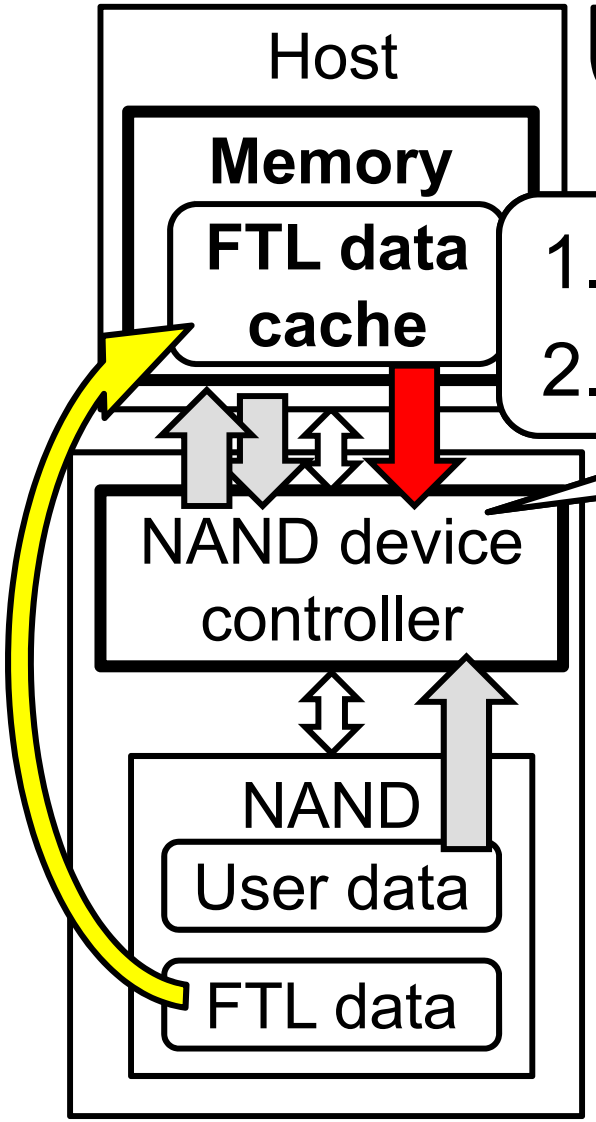
Large capacity on-chip RAM increases...

- **Chip power consumption**
- **Chip size (height)**

Critical problems for
embedded NAND storage device

Is there any alternative to on-chip RAM?

"Let Them Use Host Memory"
Unified Memory Architecture (UMA)



Makes NAND device controller available to access host memory

1. Reads **host memory** to get FTL data
2. Reads NAND to get user data

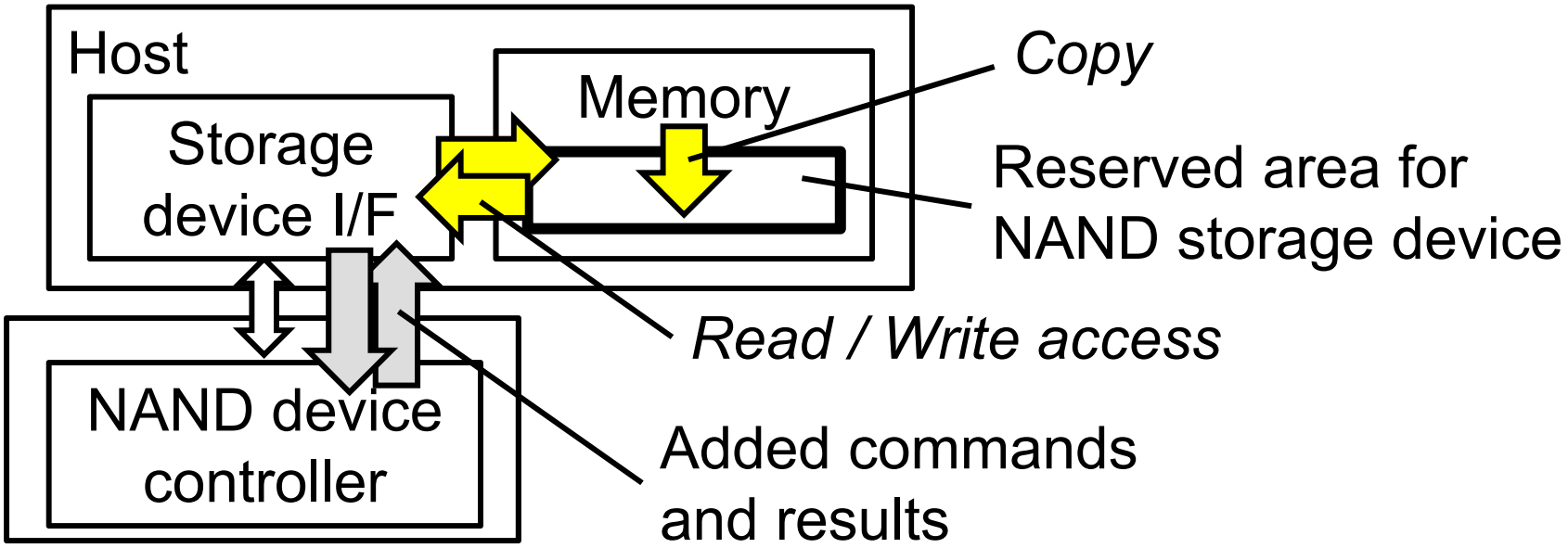
Host memory read latency
<< NAND read latency

Reduces random read latency

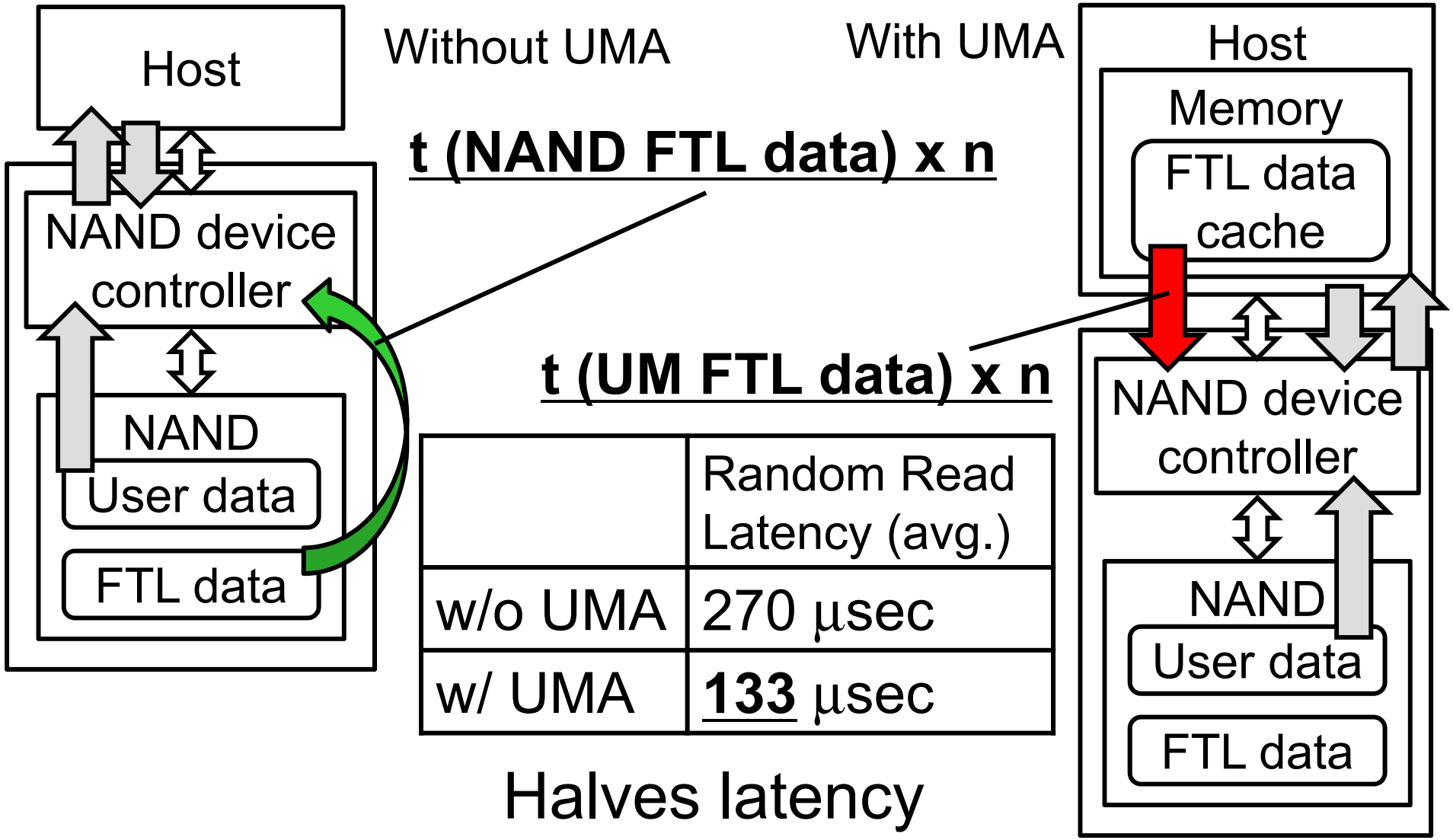
No increase in chip power consumption and chip size

What We Did to Introduce UMA

- Extended UFS standard
  - Added three commands to use host memory efficiently
    - Read / Write / Copy
- Modified NAND device controller to support extended UFS standard
  - Storage device I/F also needs to support extended UFS standard



How UMA is Effective



	Random Read Latency (avg.)
w/o UMA	270 µsec
w/ UMA	**133** µsec

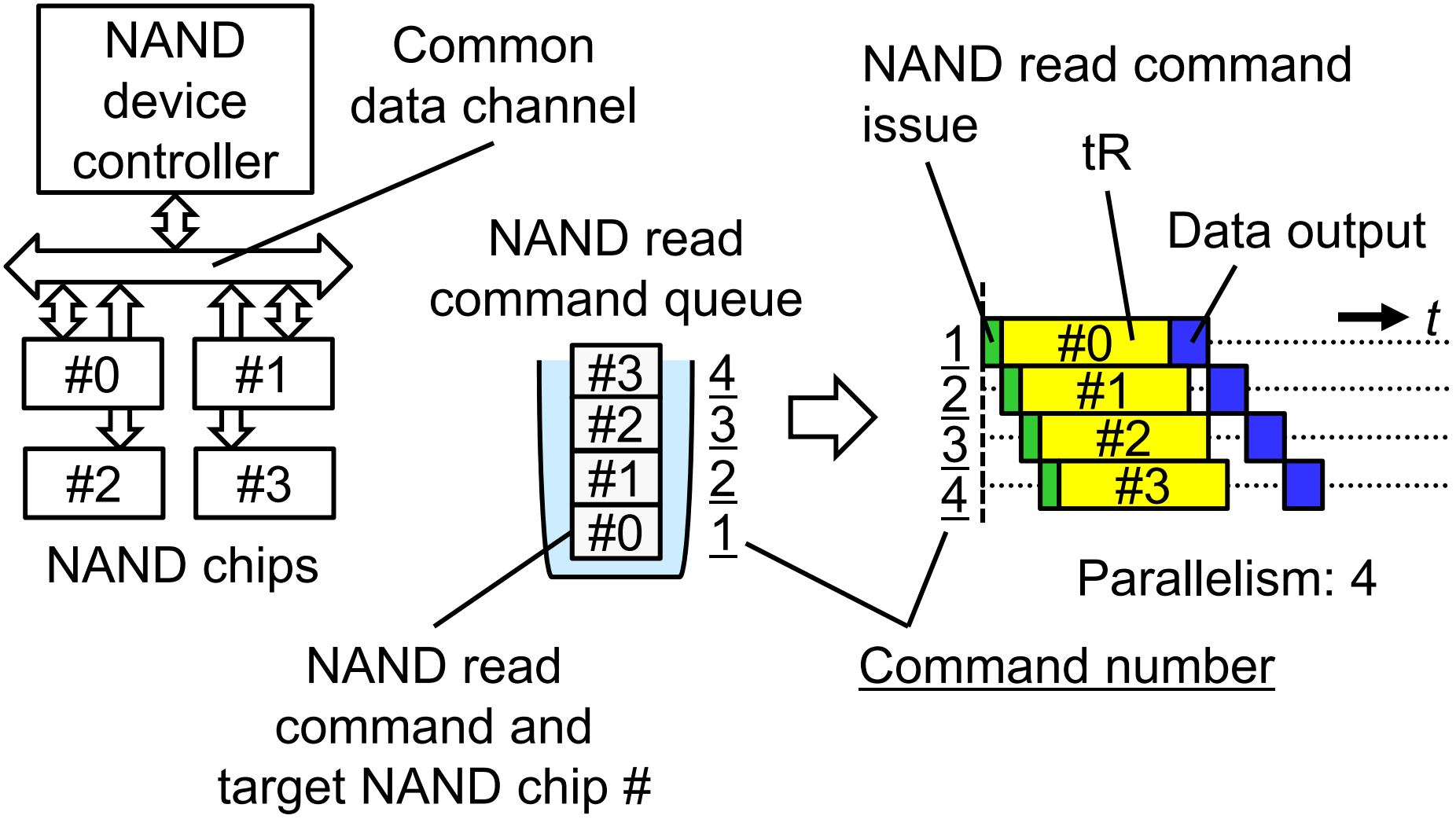
Halves latency

Works well for latency reduction

Outline

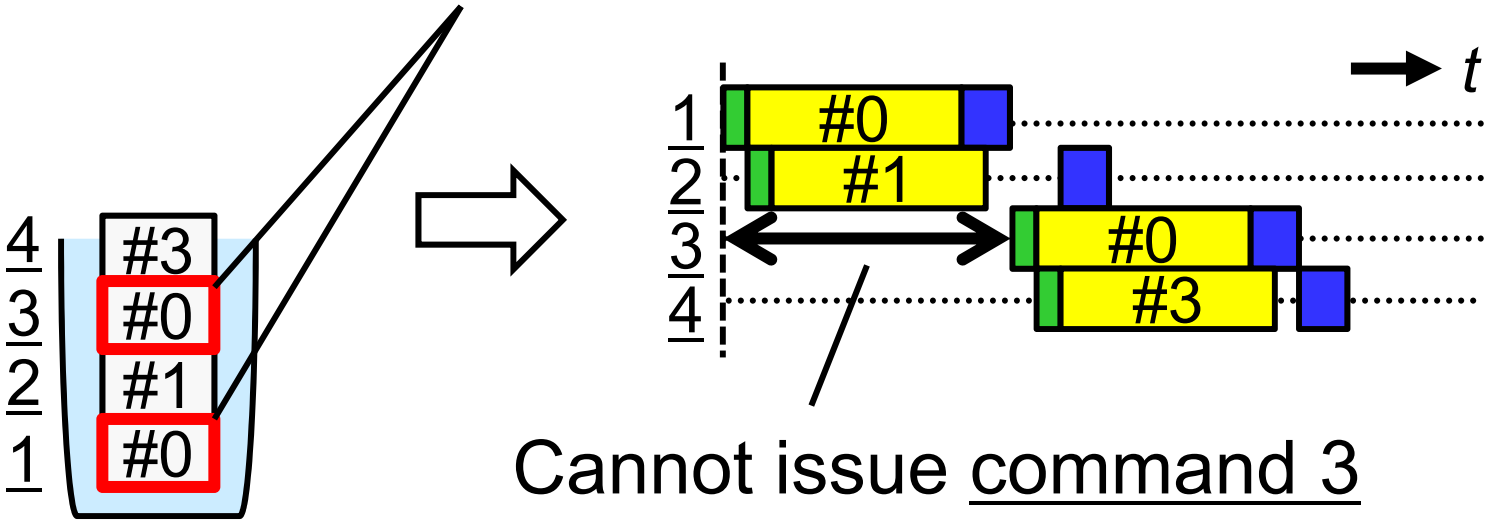
- Background and approaches
- Unified Memory Architecture
- Random Read Command Processor (RRCP)
- Synchronized-injection CDR
- Results

What's NAND Parallelism?



Increases random read throughput

Commands in Queue Usually Have Target Conflicts



Cannot issue command 3 until NAND chip #0 finishes command 1 processing

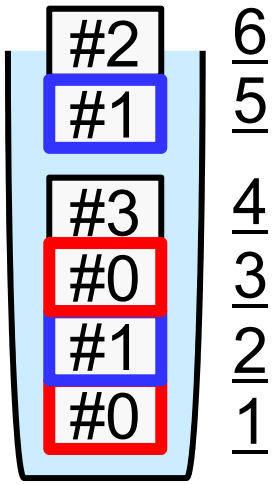
⇨ Parallelism: 2

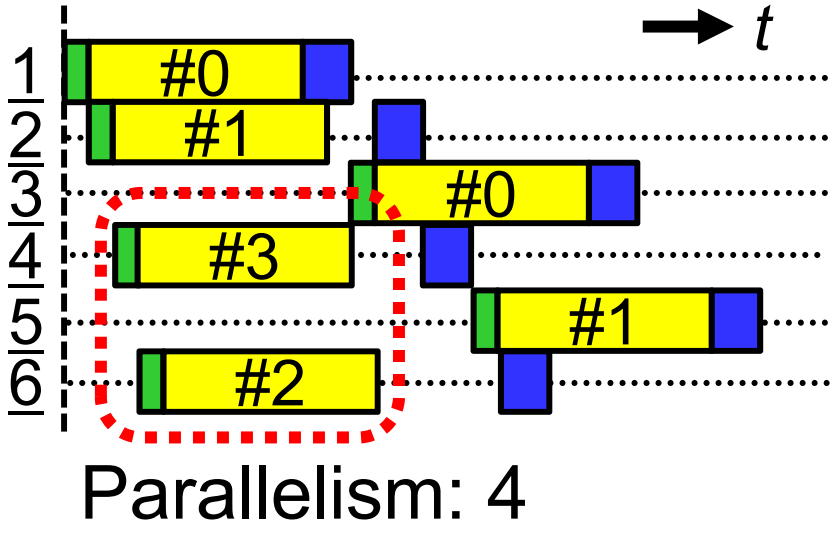
Target conflicts degrade parallelism

Reduce Impact of Target Conflicts

- **Deep command queue**
- **Out-of-order command processing**
  - Queue depth: 4 → 6
  - NAND device controller can issue commands in arbitrary order

2 target conflicts

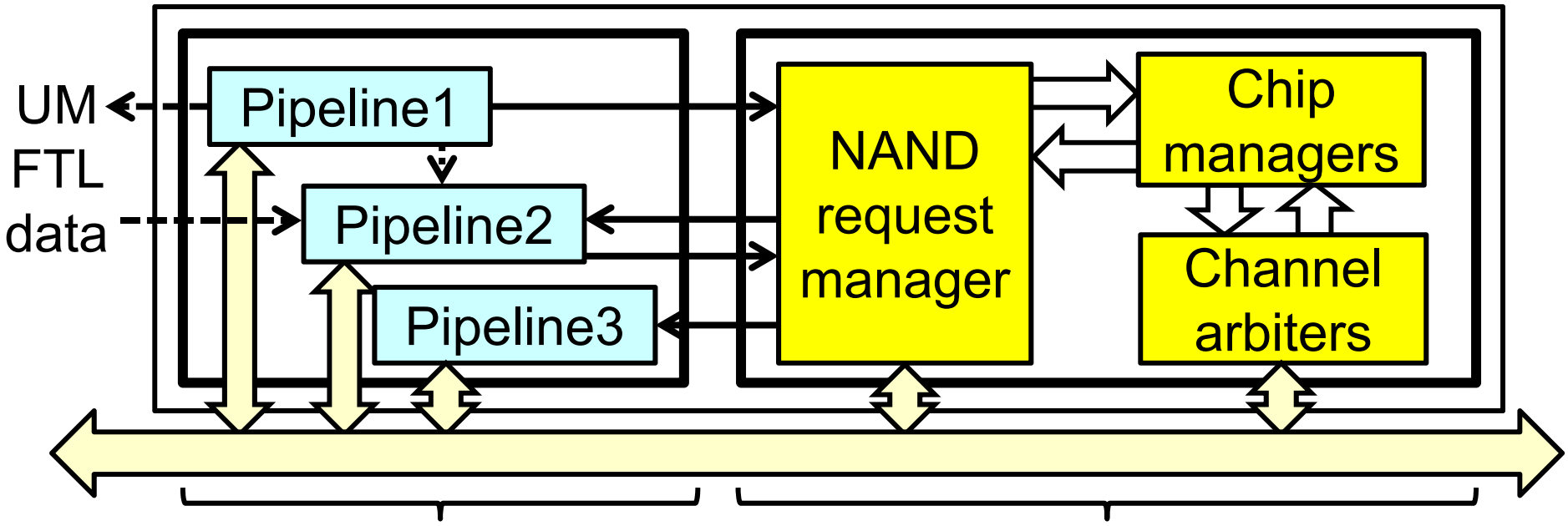




Parallelism: 4

Expected parallelism becomes higher

Random Read Command Processor

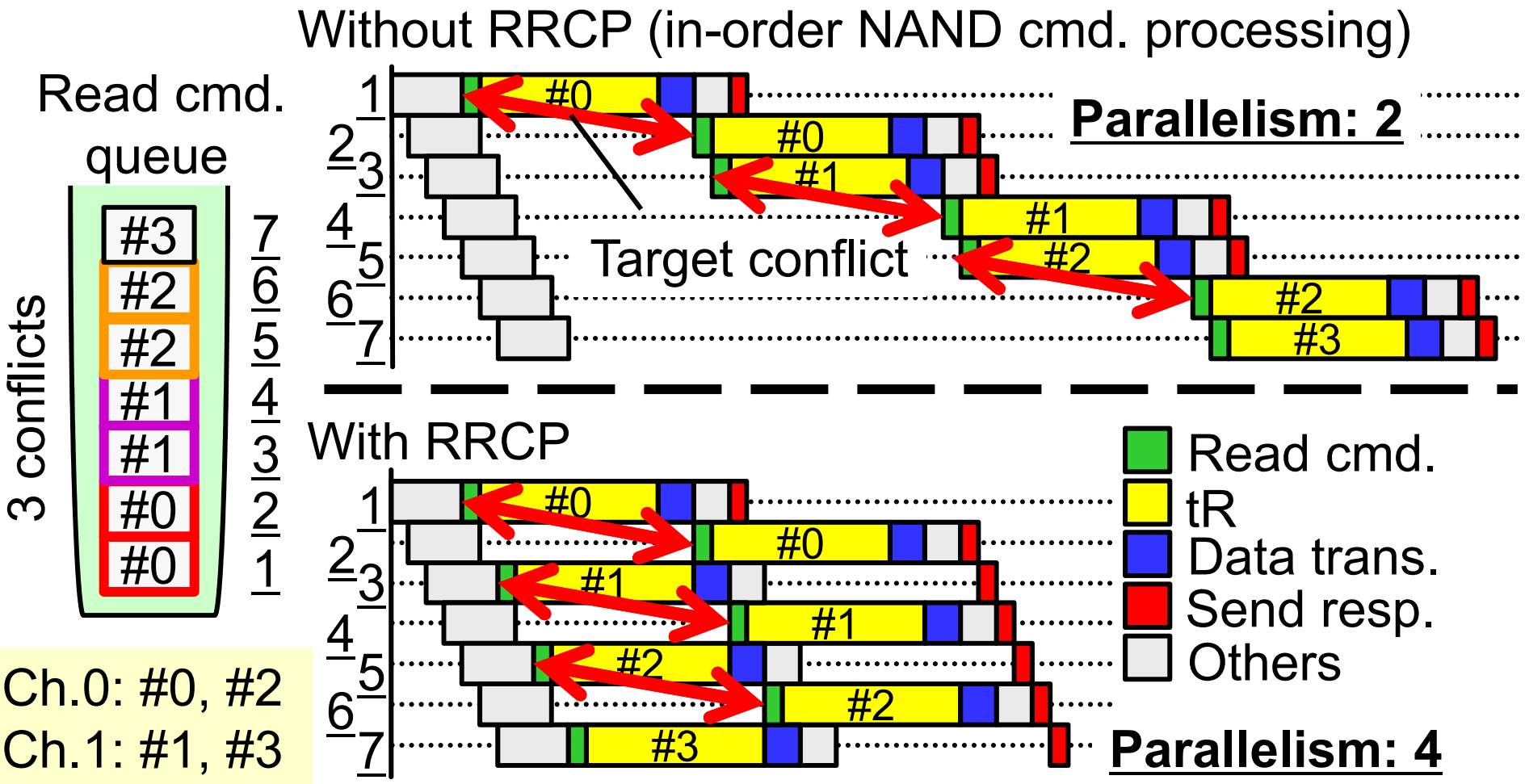


Front-end
Random read command processing

Back-end
Out-of-order NAND read command processing

Behaves as depth 16 out-of-order command queue

How RRCP is Effective



Increases expected parallelism 113%
(8 NAND chips)

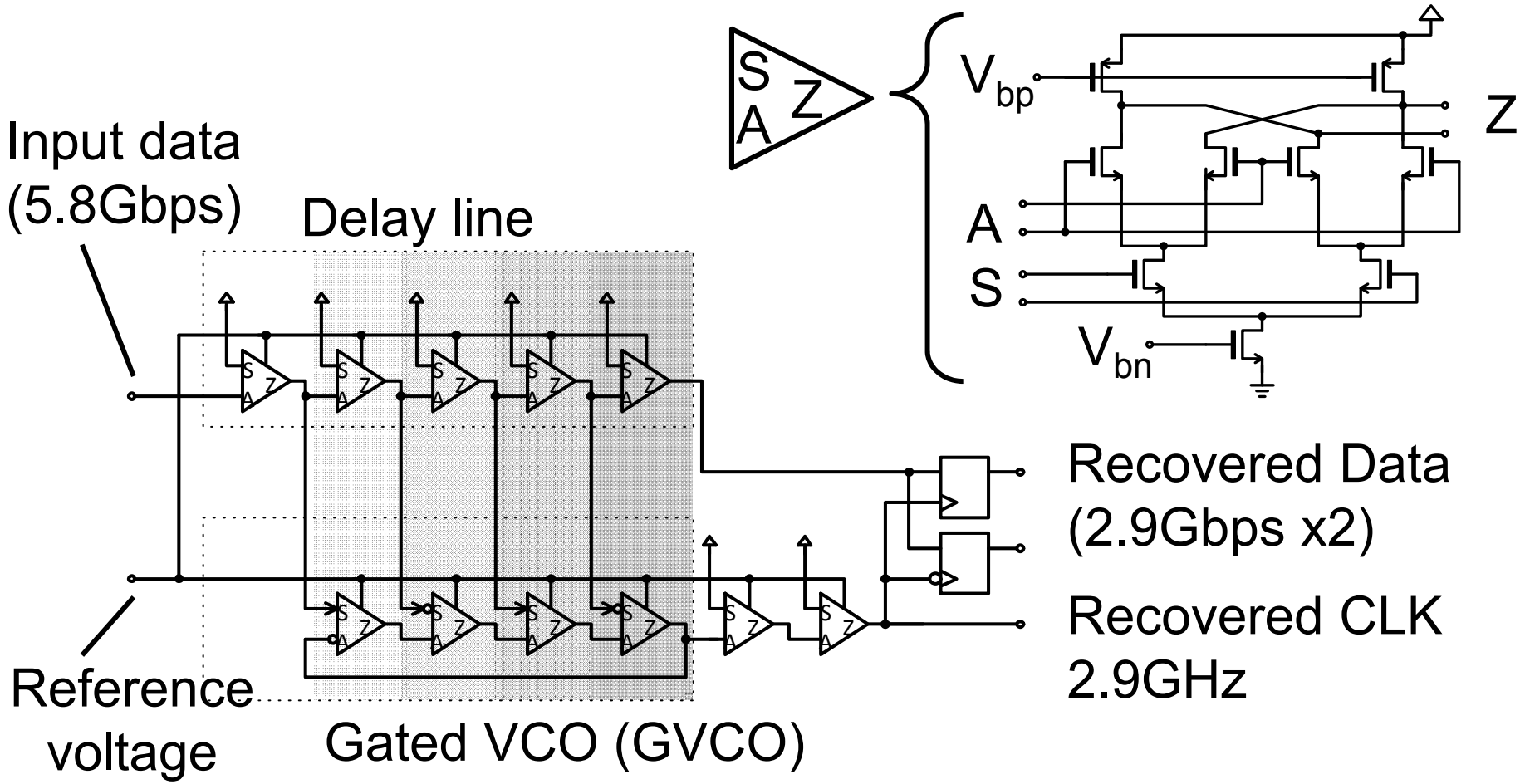
Outline

- Background and approaches
- Unified Memory Architecture
- Random Read Command Processor (RRCP)
- Synchronized-injection CDR
- Results

M-PHY Requirements and CDR

- Our M-PHY module must be
  - High speed (5.8Gbps x 2-lane)
  - Low power
  - High jitter tolerance
    - Minimum jitter tolerance is specified
- Jitter tolerance should be cared by CDR circuit
  - High speed, low power and high jitter tolerance CDR circuit was needed
- Injection-based CDR circuit
  - High speed (compared to general PLL-based CDR)
  - Low jitter tolerance

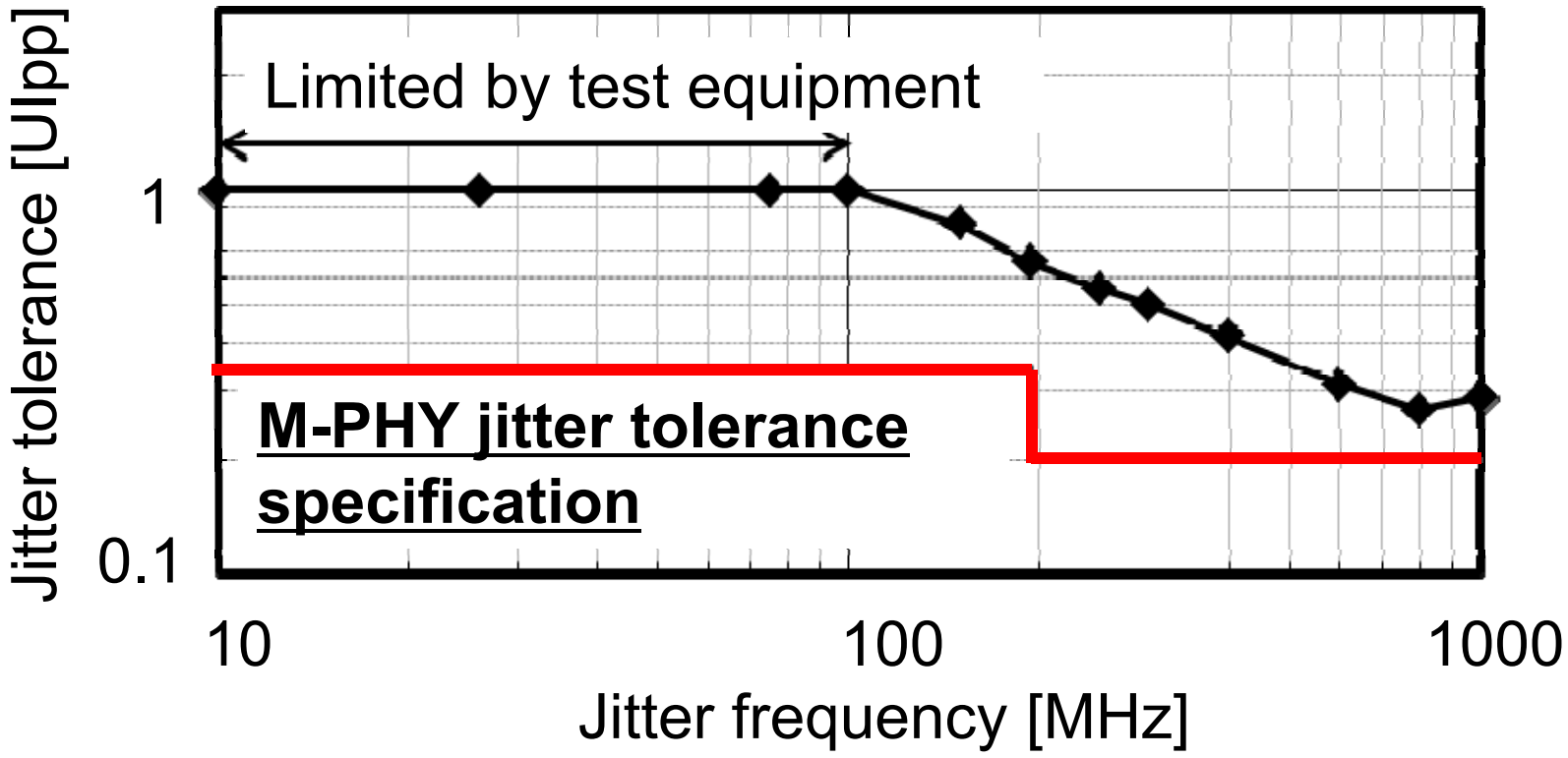
Developed Synchronized-injection CDR



Using same cells for delay line and GVCO

⇨ **High jitter tolerance**

Characteristic of Synchronized-injection CDR



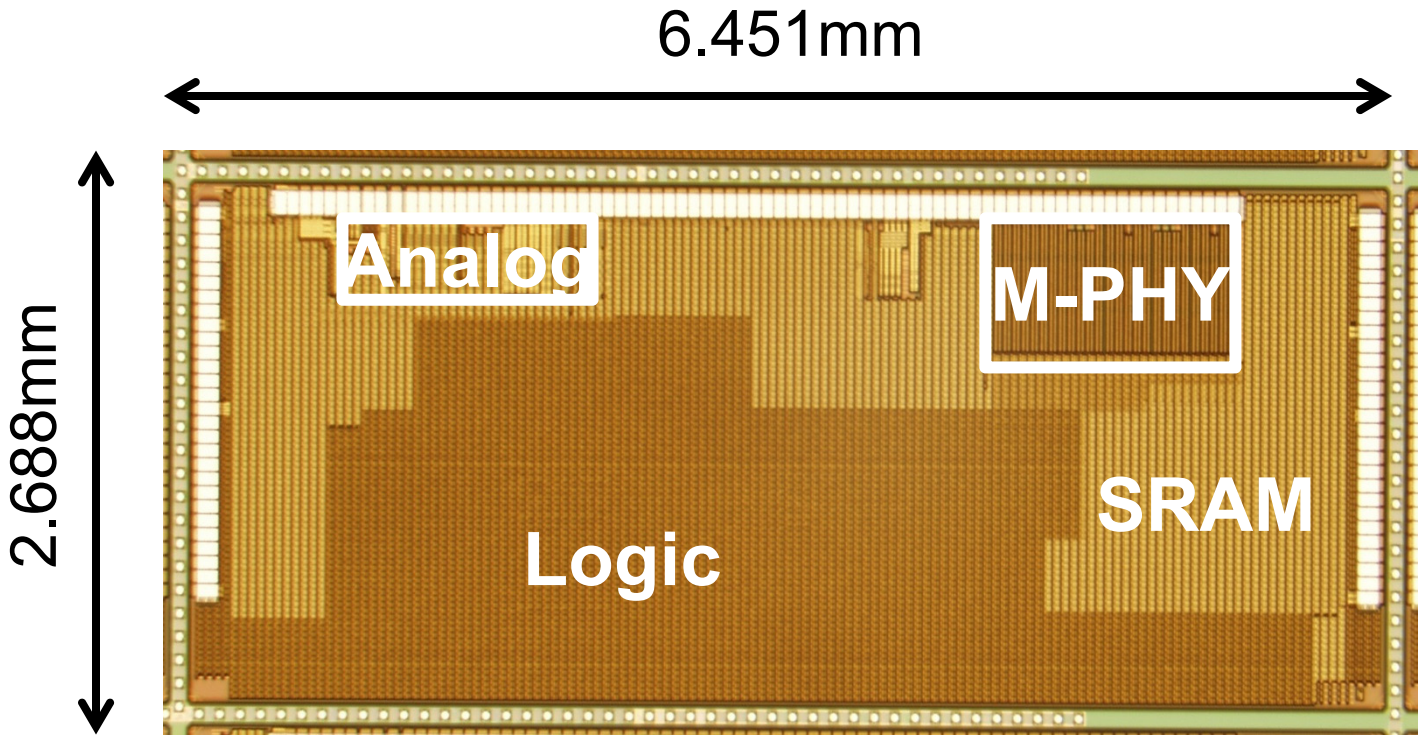
Power consumption: 7.0mW / lane

Low power / high jitter tolerance CDR

Outline

- Background and approaches
- Unified Memory Architecture
- Random Read Command Processor (RRCP)
- Synchronized-injection CDR
- Results

Micrograph of NAND Device Controller Chip



Device Performance and Features

		This work (UFS 2.0)	eMMC v5.0	SATA μSSD	SSD
4KB random read (KIOPS)	w/o UM	32.3	7.0	9.0	35-100
	w/ UM	**66.3**†			
Sequential read (MB/s)		**690**	250	450	500-550

† Depends on UM size, access range and host latency

Process		40nm CMOS
Controller chip size		17.34 mm^2
Package size (with NAND chips)		**11.5mm x 13.0mm** x 1.2mm
Package power	Active (Seq. write / read)	< **1.5 W**‡ / < **1.5 W**‡
	Idle / Sleep	< 1.45 mW‡ / < 0.30 mW

‡ Depends on storage capacity

Conclusion

- Developed UFS 2.0 embedded NAND storage device
- Improved random read performance w/ new features
  - Unified memory architecture
  - Random read command processor
- Accomplished desired higher read performance
  - Random read: **66.3KIOPS**
  - Sequential read: **690MB/s**
- With typical size / power consumption
  - Size (with NAND chips) : **11.5mm x 13.0mm** x 1.2mm
  - Active (seq. read / write) : < **1.5 W**

Embedded 1Mb ReRAM in 28nm CMOS with 0.27-to-1V Read Swing-Sample-and-Couple Sense Amplifier and Self-Boost-Write-Termination Scheme

Meng-Fan Chang[1], Jui-Jen Wu[1], Tun-Fei Chien[1], Yen-Chen Liu[1], Ting-Chin Yang[1], Wen-Chao Shen[1], Ya-Chin King[1], Chorng-Jung Lin[1], Ku-Feng Lin[2], Yu-Der Chih[2], Sreedhar Natarajan[2], and Jonathan Chang[2]

[1]National Tsing Hua University (NTHU), Hsinchu, Taiwan

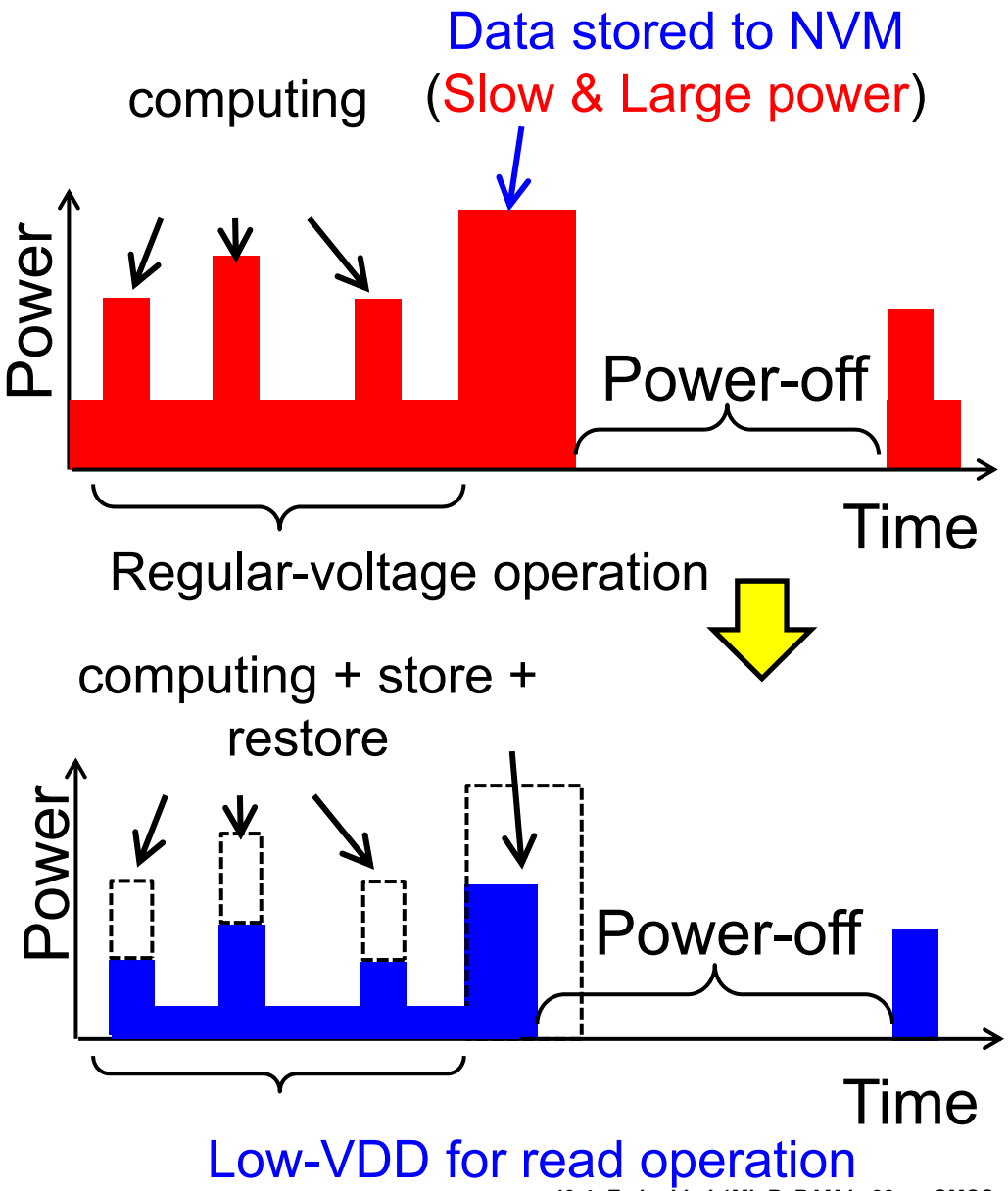
[2] TSMC, Hsinchu, Taiwan

ISSCC 2014 – Session 19.4

Outline

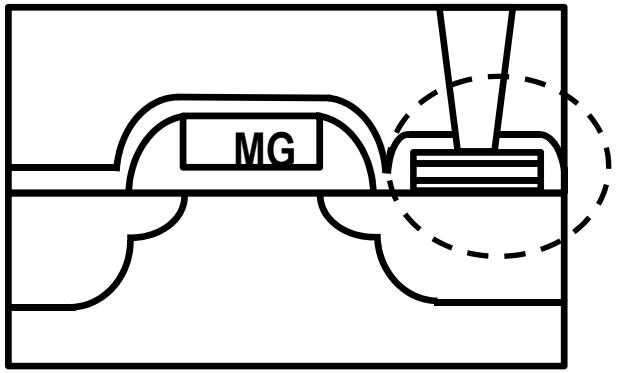
- Introduction
- Swing-Sample-and-Couple Voltage Mode Sense Amplifier (SSC-VSA)
- Self-Boost-Write-Termination Scheme (SBWT)
- Measurement Results
- Conclusion

NVM in Energy-Efficient Systems

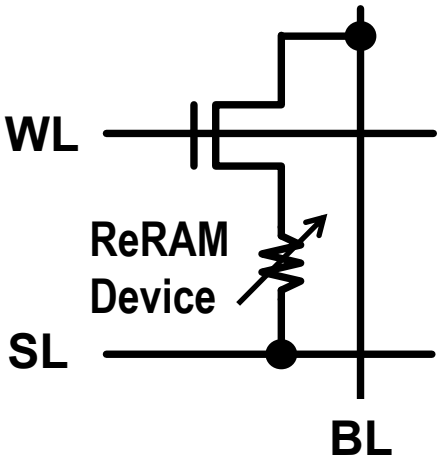


- NVM enable power-off
  - Code/data storage
  - Reduce standby power
- Low-Energy ReRAM
  - Low write power
  - Low write voltage
  - Seldom-write (power off)
    - Use nominal VDD
  - Frequent-read (code)
    - Use low VDD

Logic-Compatible ReRAM Device



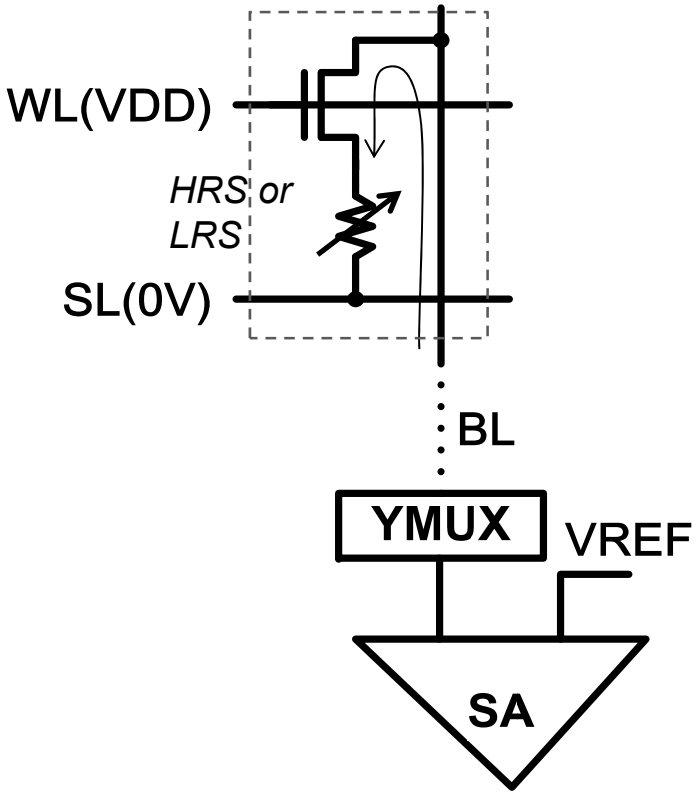
NTHU/TSMC, IEDM 2012

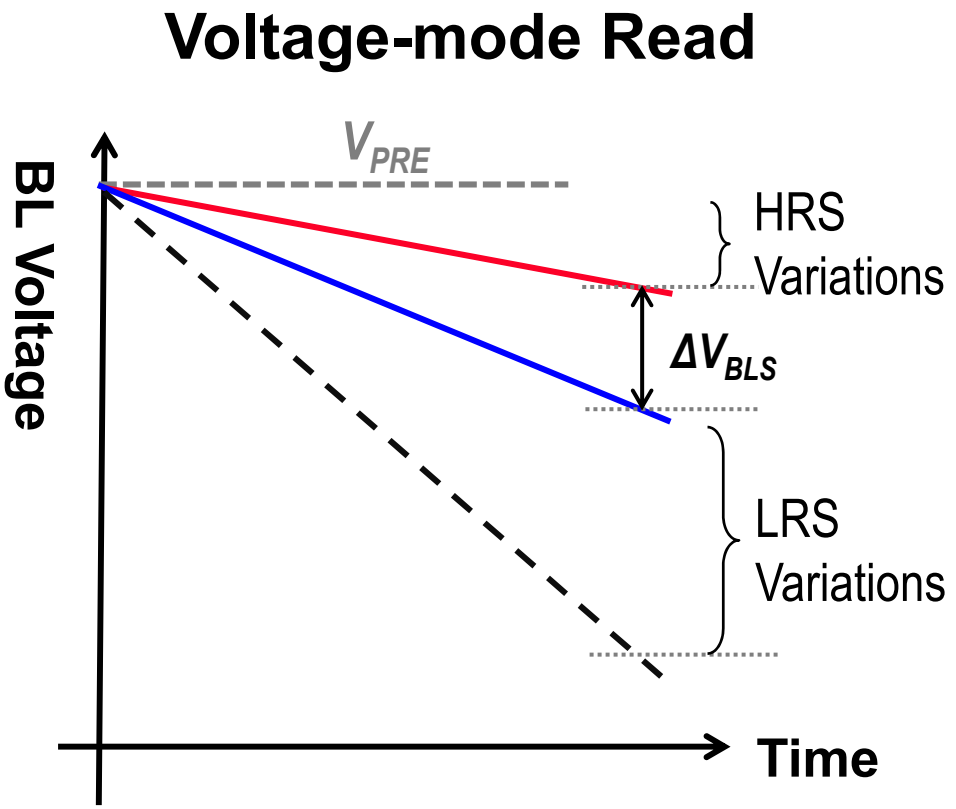


	SET	RESET
WL	$V_{WL\text{-}SET}$	$V_{WL\text{-}RESET}$
BL	0V	0V
SL	V_{SET}	V_{RESET}
Res.	HRS to LRS	LRS to HRS
Speed	T_{SET}	T_{RESET}

- Highly Compatible with CMOS BEOL
- Low write current and compact area

Challenges of ReRAM - Read

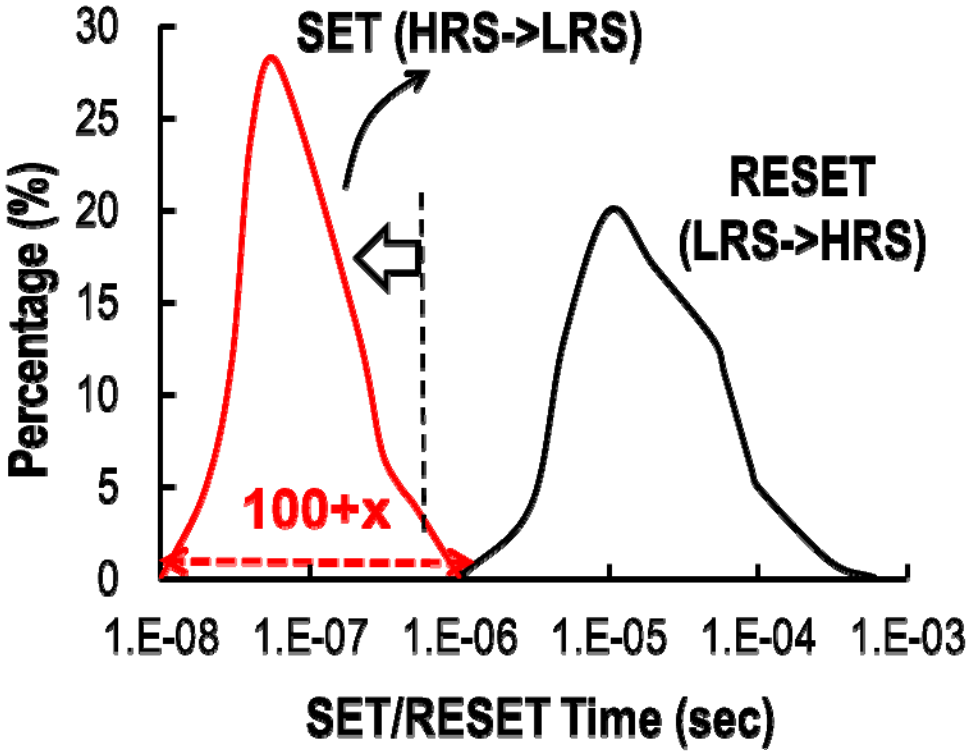


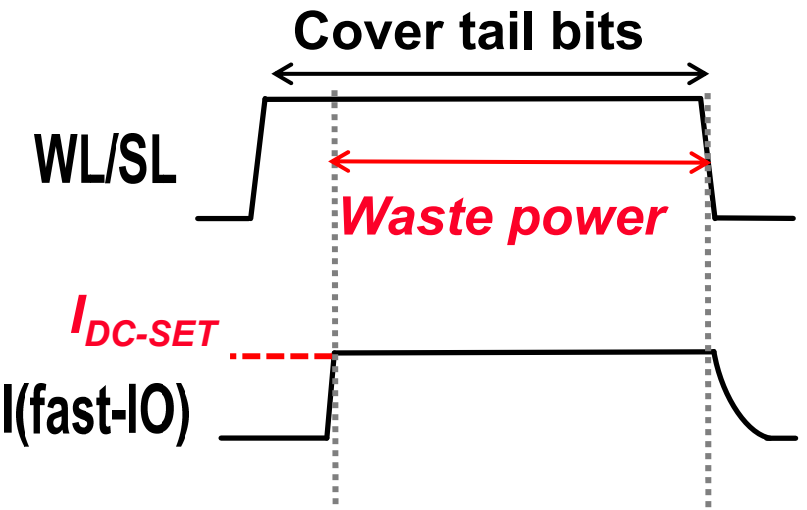


- Small R-ratio + High LRS resistance (R_L)
  - Both HRS and LRS cell have read BL swings (V_{BLS})
  - Small ΔV_{BLS} between tail bits, especially at low-VDD

=> Limited read VDDmin, slow read speed

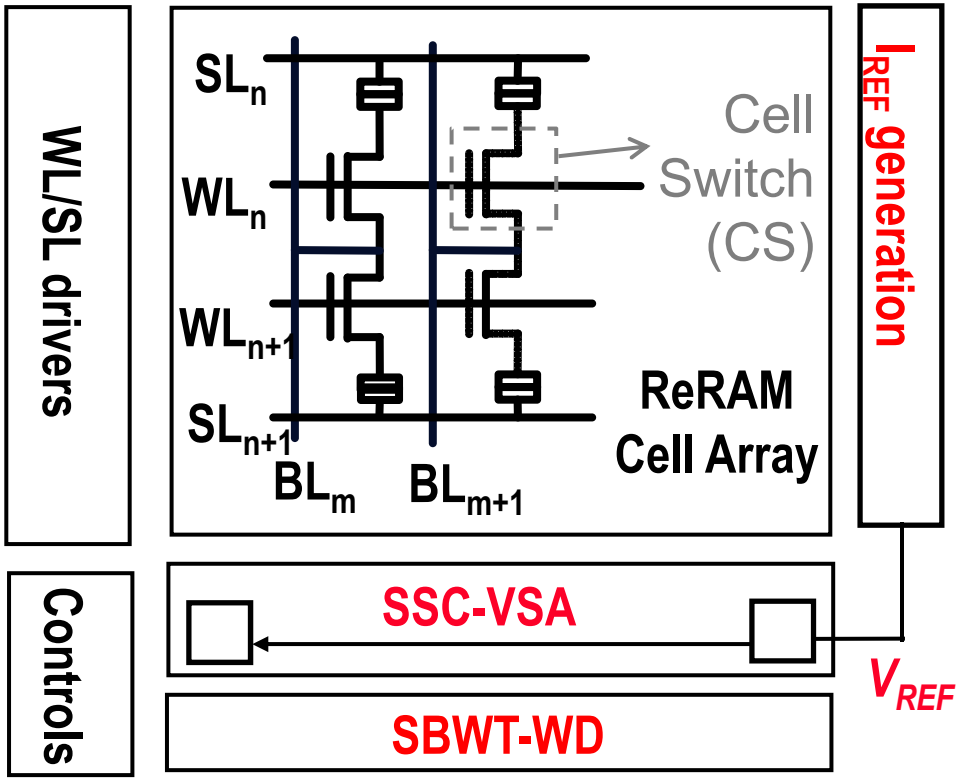
Challenges of ReRAM - Write





- Wide distribution in SET/RESET time
- A fast SET operation causes large DC current (I_{DC_SET})
  - WL/SL are kept for long time to cover worse case.
  - Waste power consumption

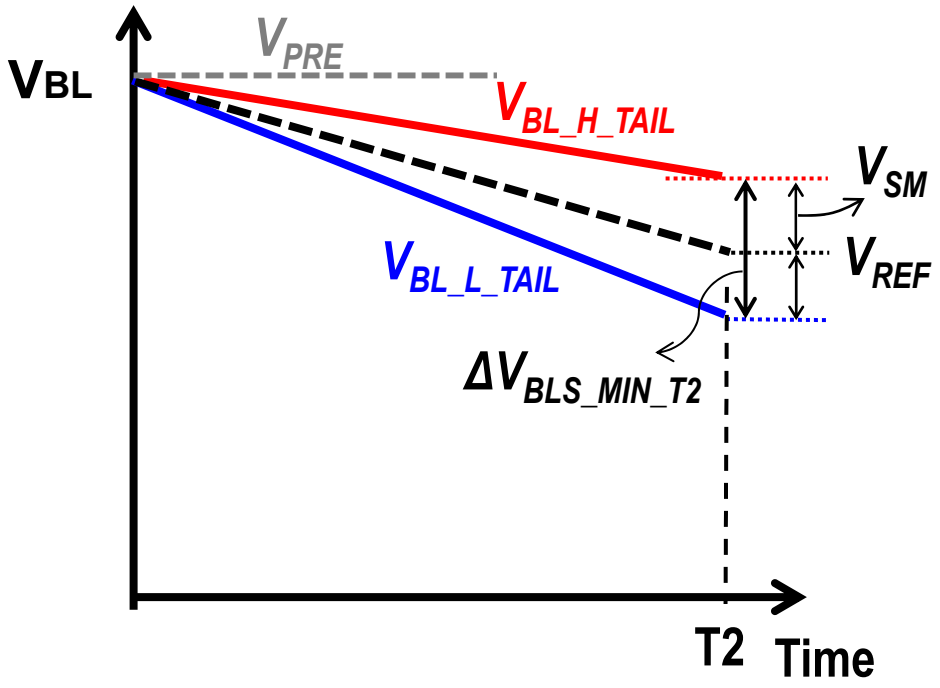
Proposed ReRAM Macro



- Low VDDmin read
  - Swing-Sample-and-Couple Voltage Mode Sense Amplifier
  - I_{REF} (V_{REF}) generation
- SET energy reduction
  - Self-Boost-Write-Termination Write-Driver

Concept of SSC-VSA

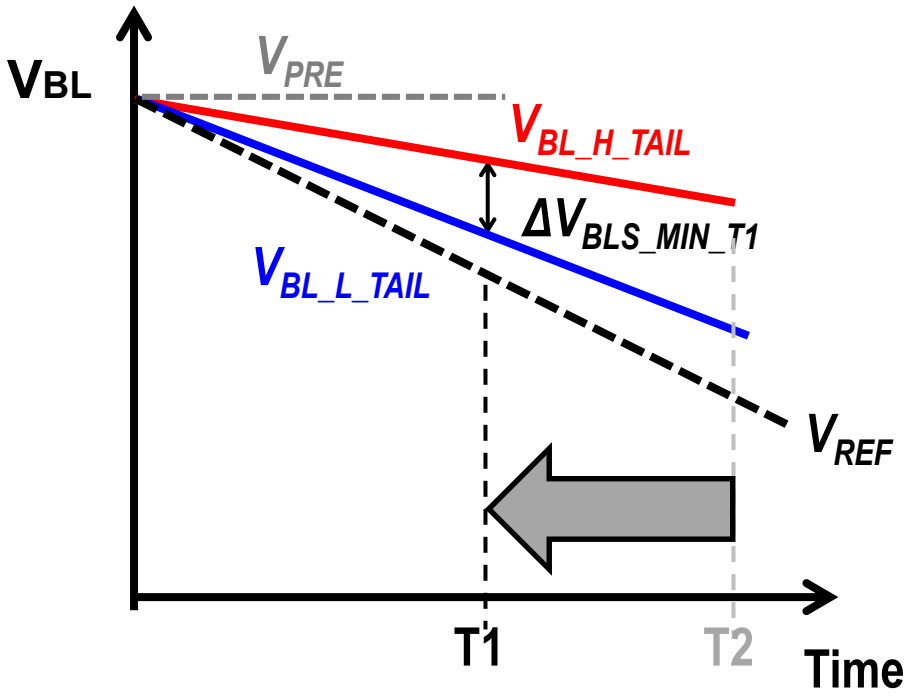
Conventional VSA



- Place V_{REF} between HRS (Read-1) and LRS (Read-0)

→ ***SM*= 0.5x ($\Delta V_{BLS\_MIN\_T2}$)**

SSC-VSA

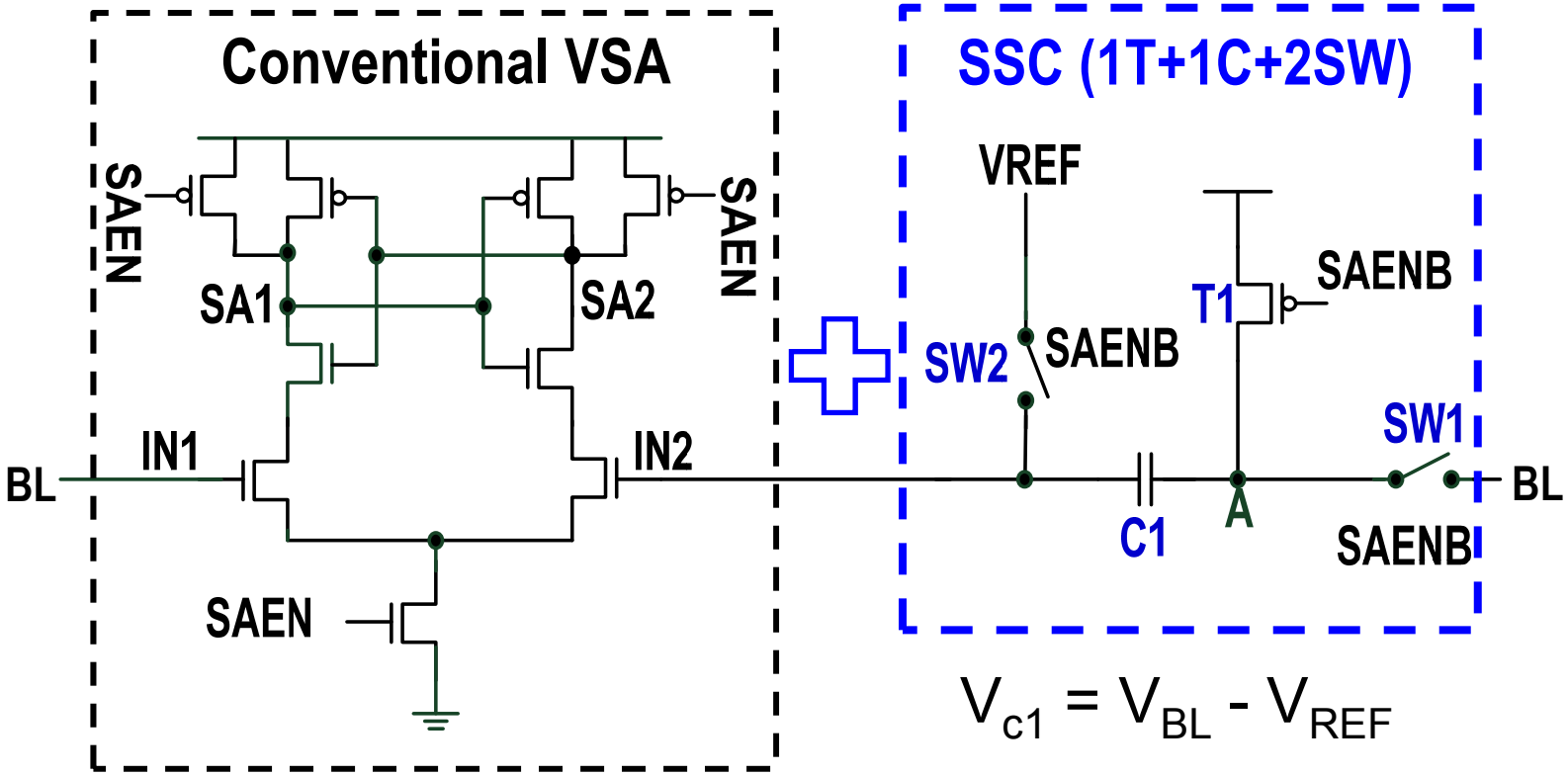


- Place V_{REF} below $V_{BL\_L\_TAIL}$
- Fully utilize ΔV_{BLS_MIN}

→ ***SM*= $\Delta V_{BLS\_MIN\_T1}$**

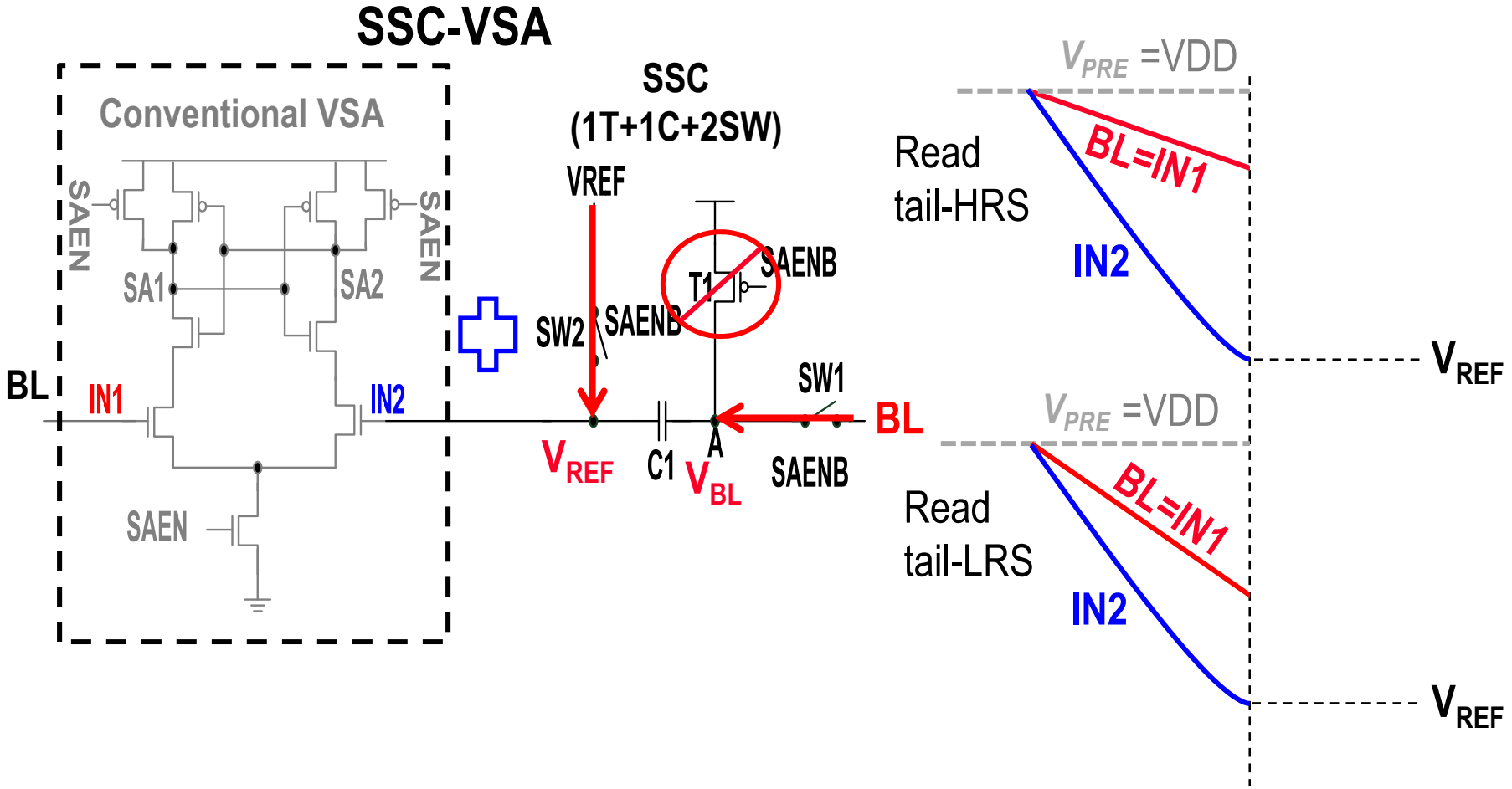
SM: sensing margin

Structure of SSC-VSA



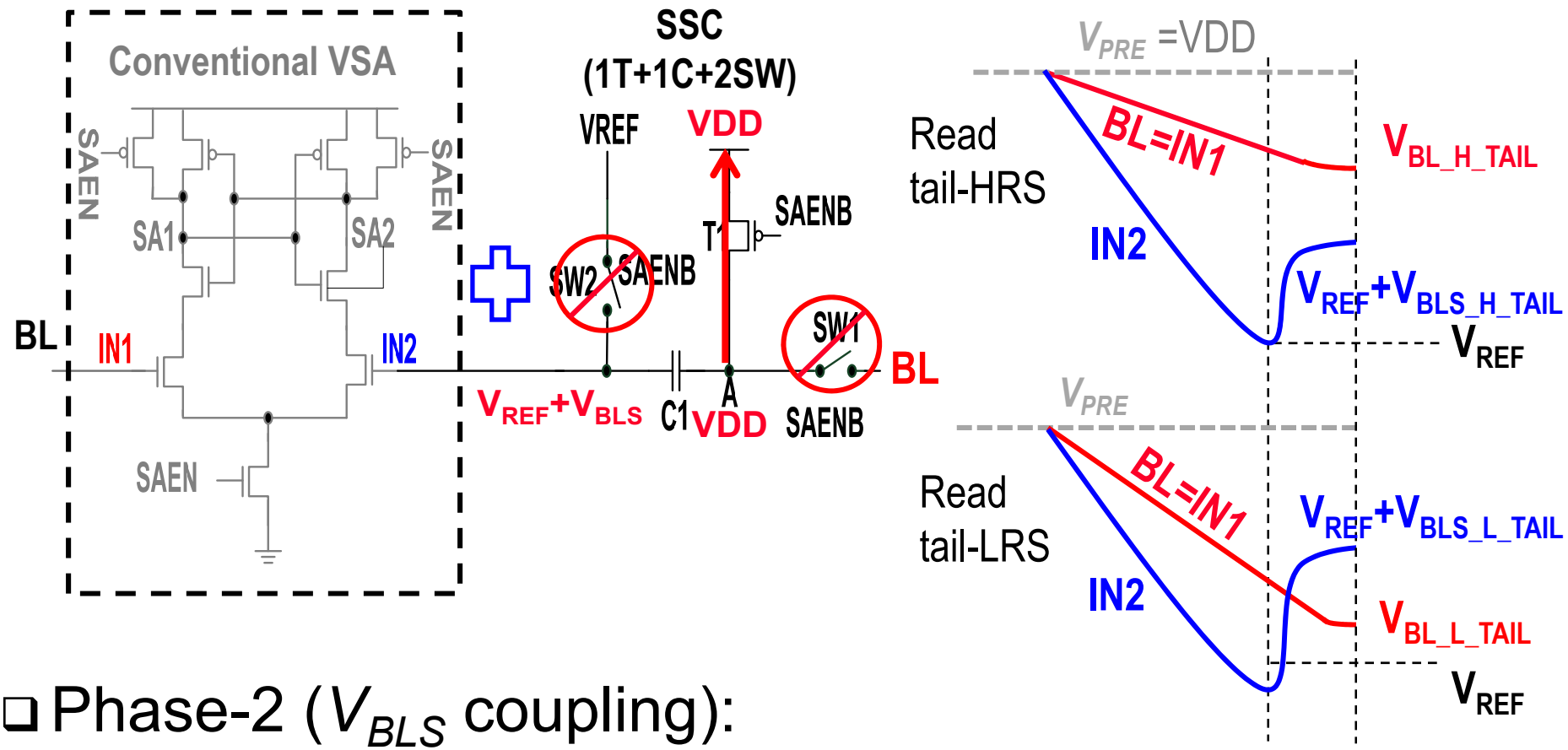
- SSC-VSA:
  - VSA: Conventional or small-offset VSA
  - SSC: 1T (PMOS) + 1C + 2 switches (SW)
- Reference voltage (V_{REF}) is generated by replica BLs

Operation of SSC-VSA - 1



- Phase-1 (V_{BLS} sampling):
  - V_{IN2} =V_{REF} , V_{IN1}=V_{BL}= VDD- V_{BLS}

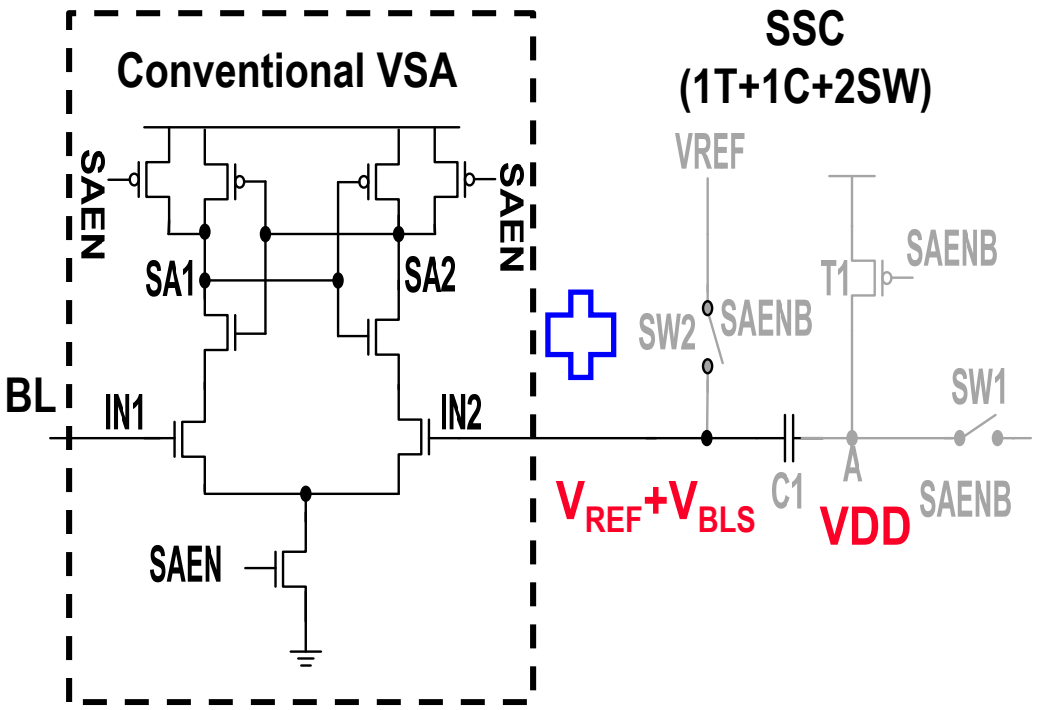
Operation of SSC-VSA - 2

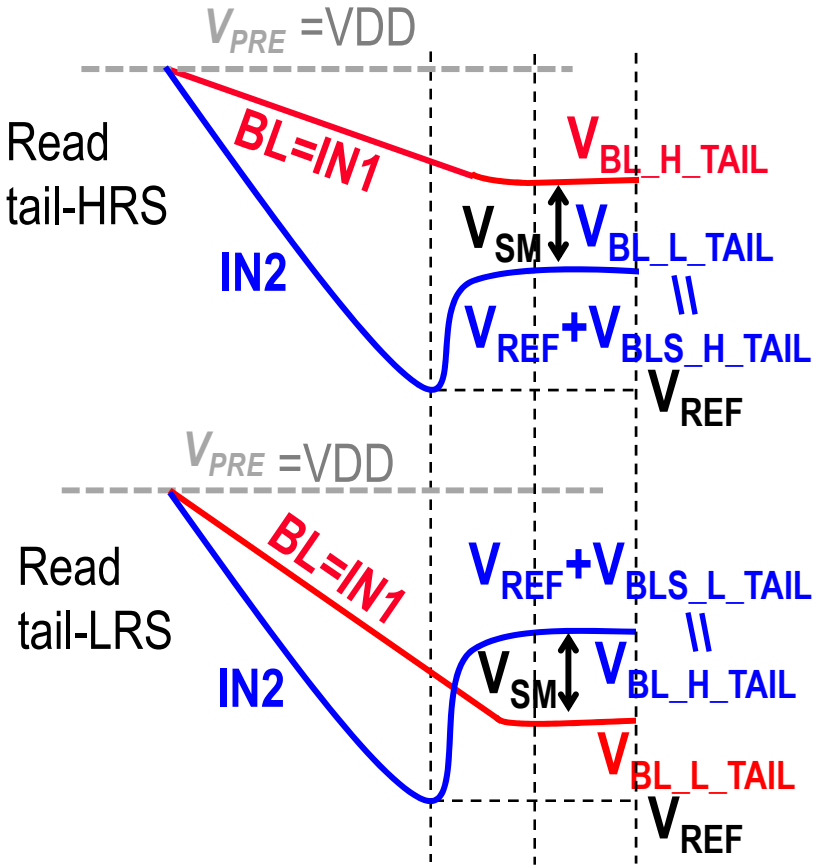


- Phase-2 (V_{BLS} coupling):
  - SW1 & SW2 are off=> nodes A and B are floating
  - SAENB=0 => pull node-A from V_{BL} to VDD

  => Boosts V_{IN2} from V_{REF} to (V_{REF}+V_{BLS})

Operation of SSC-VSA - 3

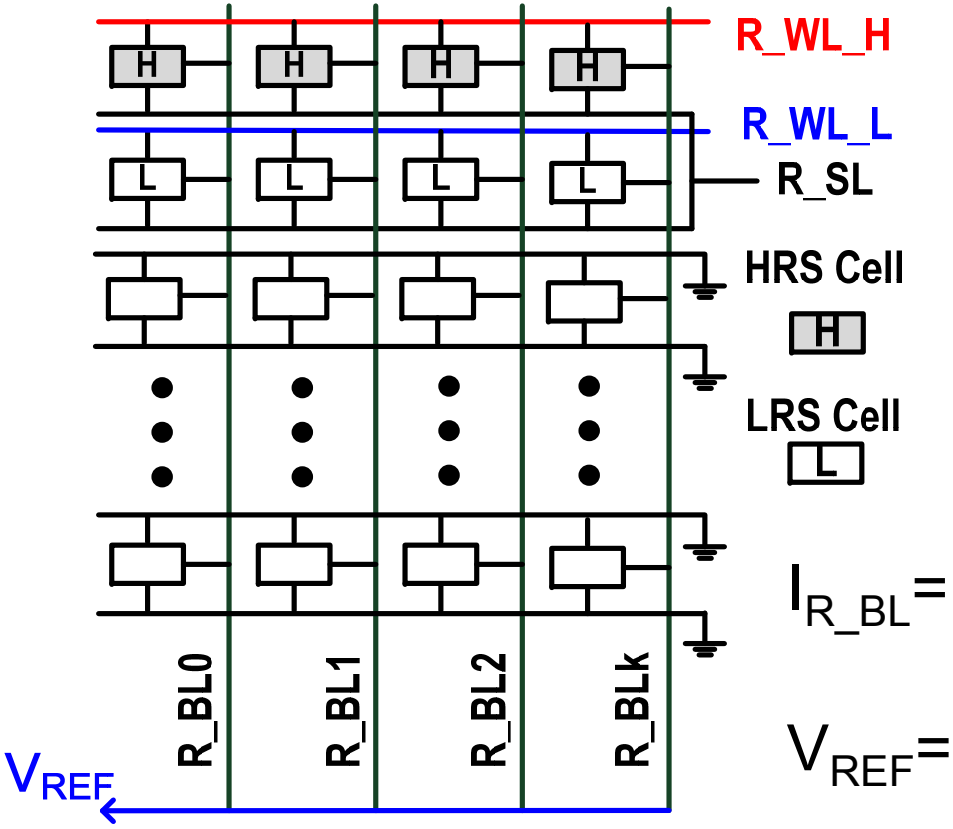




$V_{REF}=VDD-(V_{BLS\text{-}H\_TAIL}-V_{BLS\text{-}L\_TAIL})$

- Phase-3 (Comparison):
  - SAEN turns on VSA and detects $V_{IN2}-V_{IN1} = (V_{SM})$
  - $V_{SM}= V_{IN2}-V_{IN1} =2V_{BLS}-(V_{BLS\text{-}H\_TAIL}+V_{BLS\text{-}L\_TAIL})= |\Delta V_{BLS\_MIN}|$

Reference Voltage for SSC-CSA



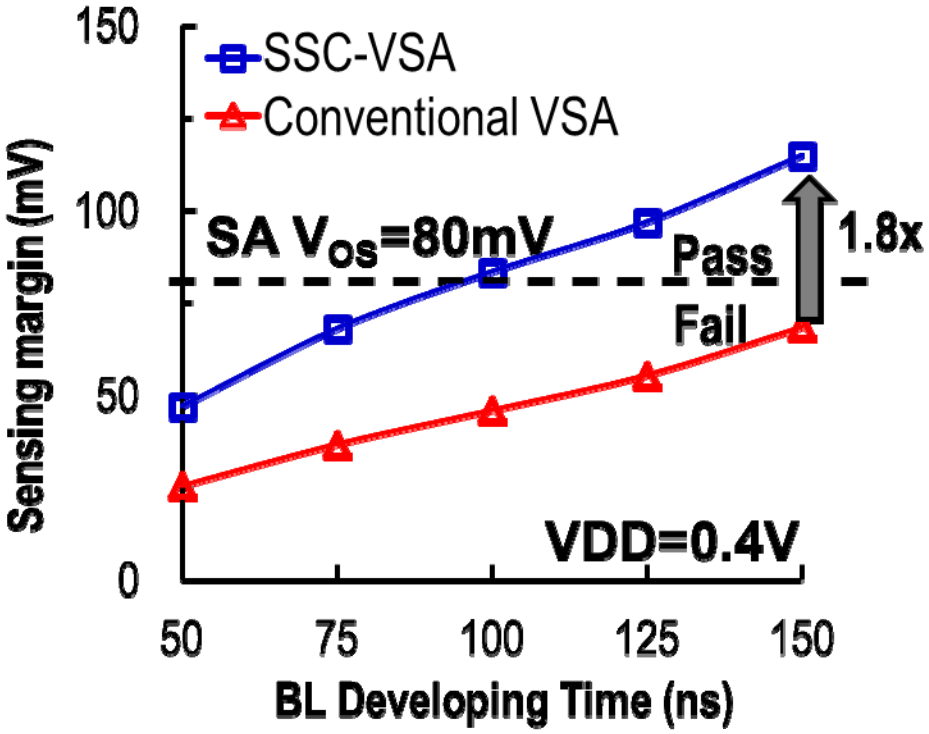
$$I_{R\_BL} = I_{HRS} + I_{LRS}$$

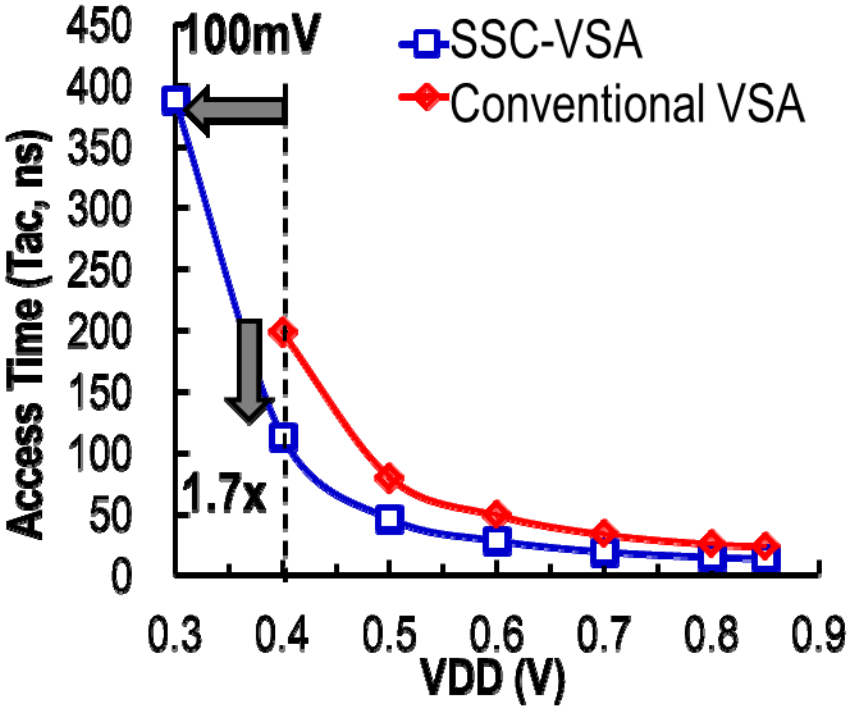
$$V_{REF} = VDD-(V_{BLS\_H\_TAIL}-V_{BLS\_L\_TAIL})$$

- Track sum of V_{BLS} of tail-HRS and tail-LRS cell
  - Averaging V_{REF} from k R_BL
- Resistance of dummy HRS/LRS cells
  - Using MLC-like operation to program cells

Performance of SSC-VSA

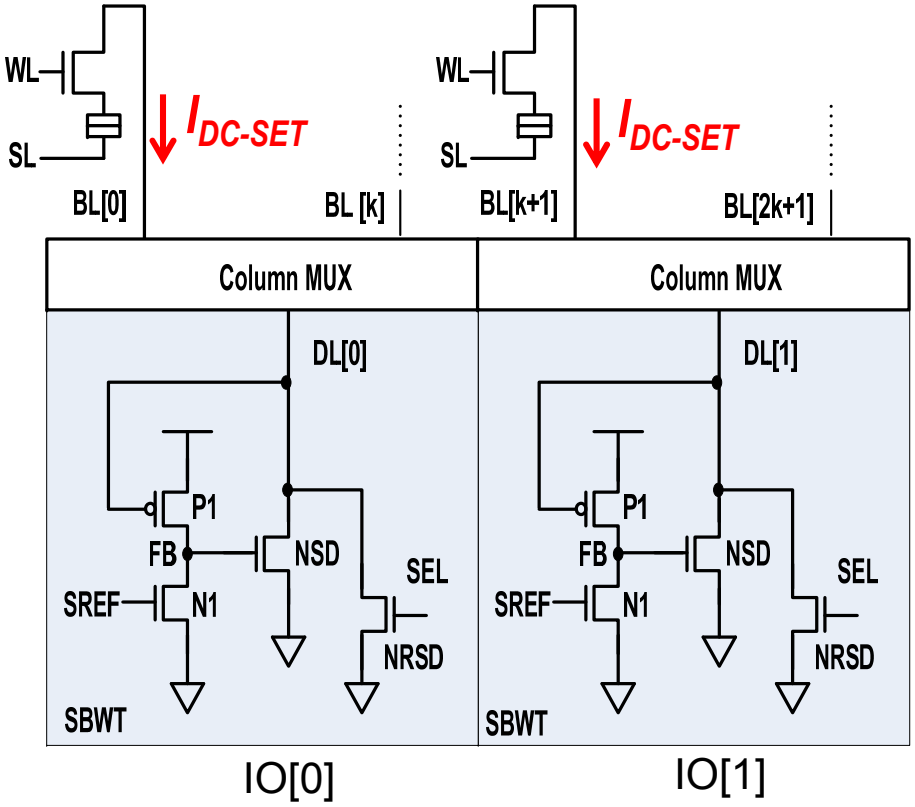
BL length=512 (at R-Ratio=5)

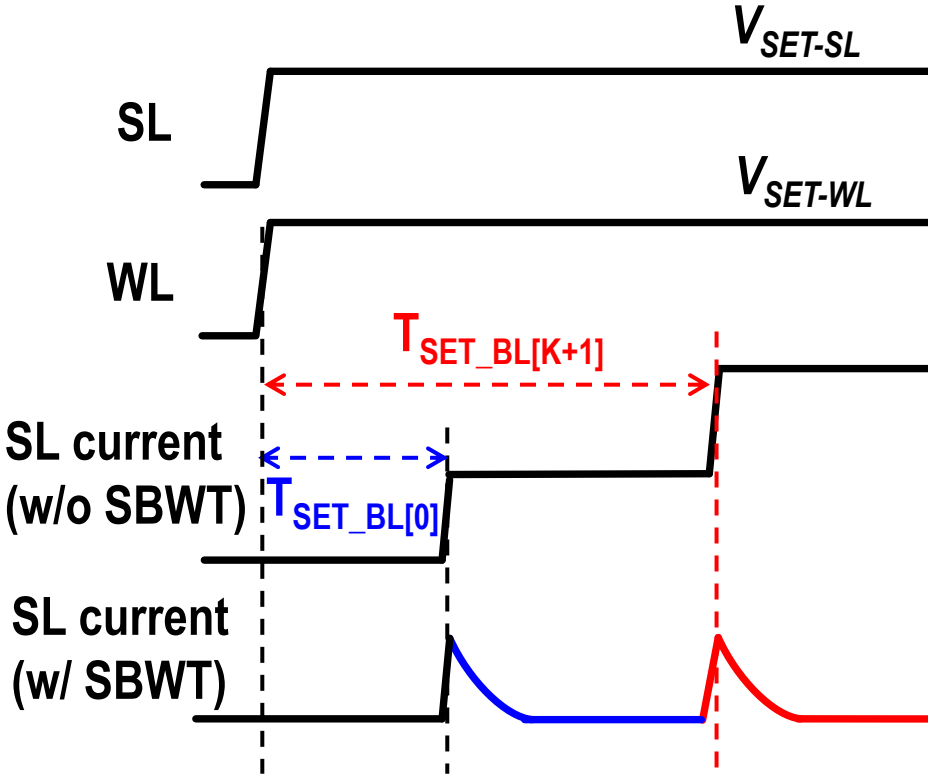




- 1.8x~2x larger SM than CD-VSA
- 100+mV lower VDDmin improvement
- 1.7+x faster access time (T_{AC})

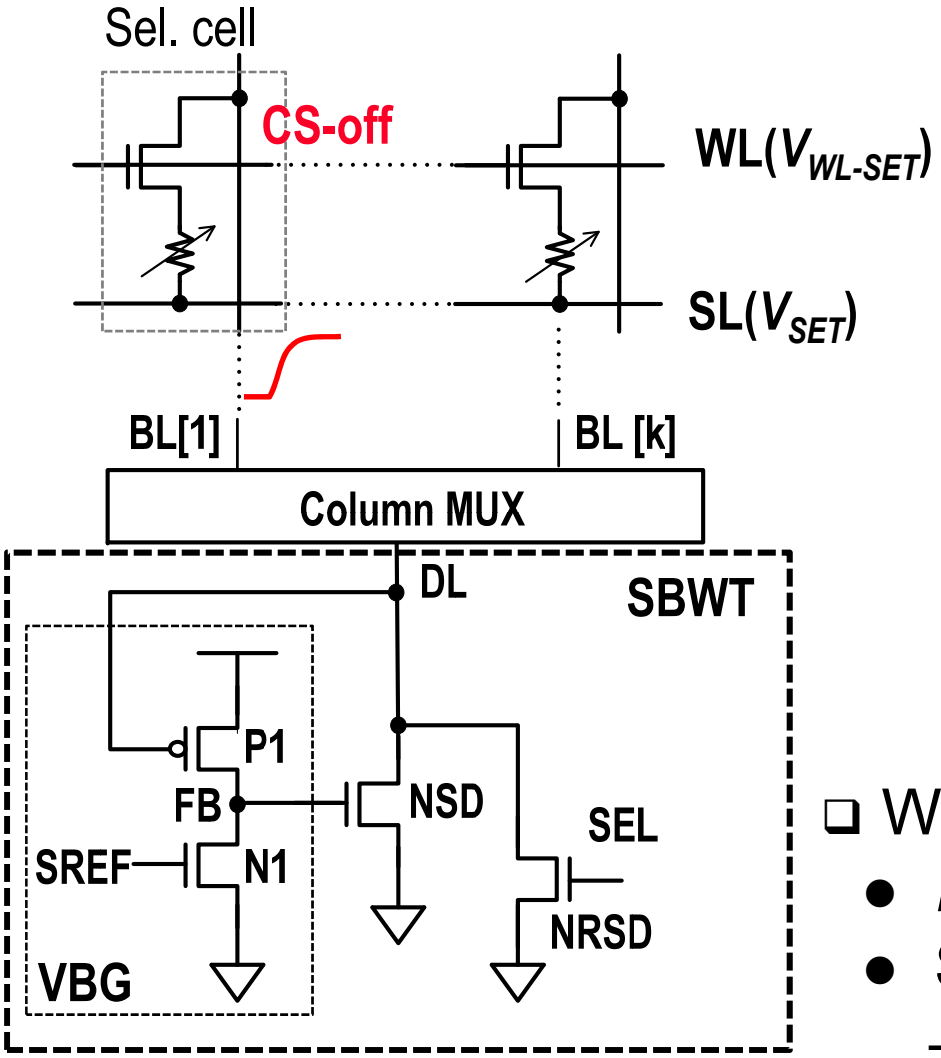
Concept and Structure of SBWT

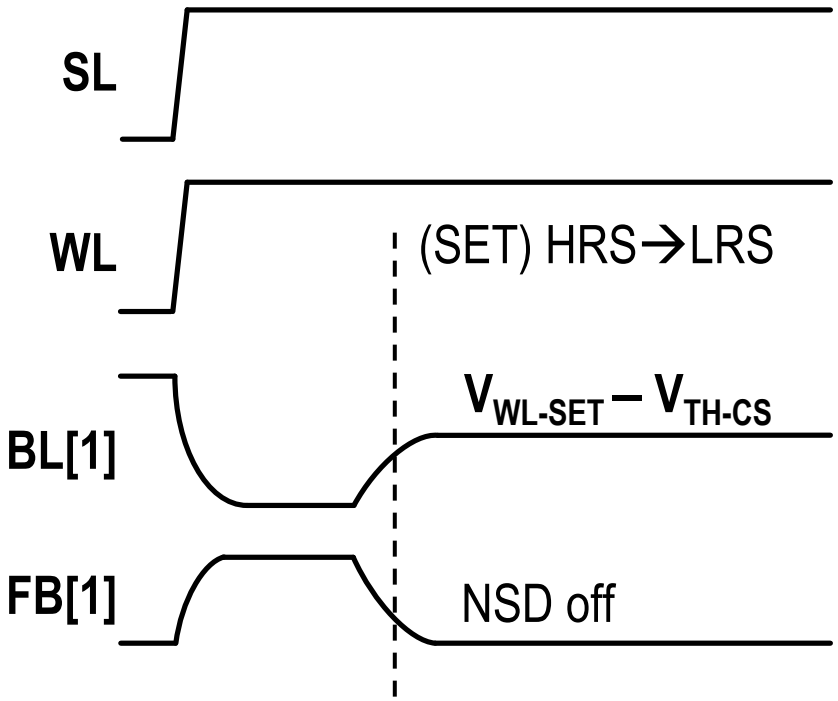




- Cut-off DC current for LRS cells
- 4T write-driver
  - 2T voltage-bias generator (VBG)
  - 1T SET driver (NSD) + 1T RESET driver (NRSD)

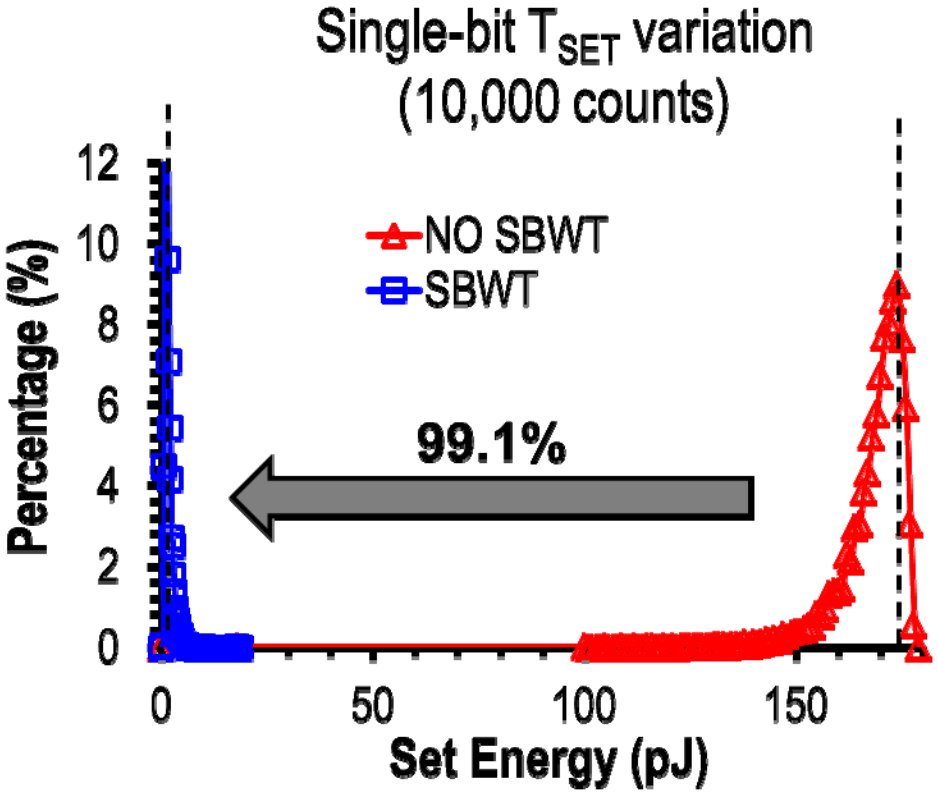
Operation of SBWT

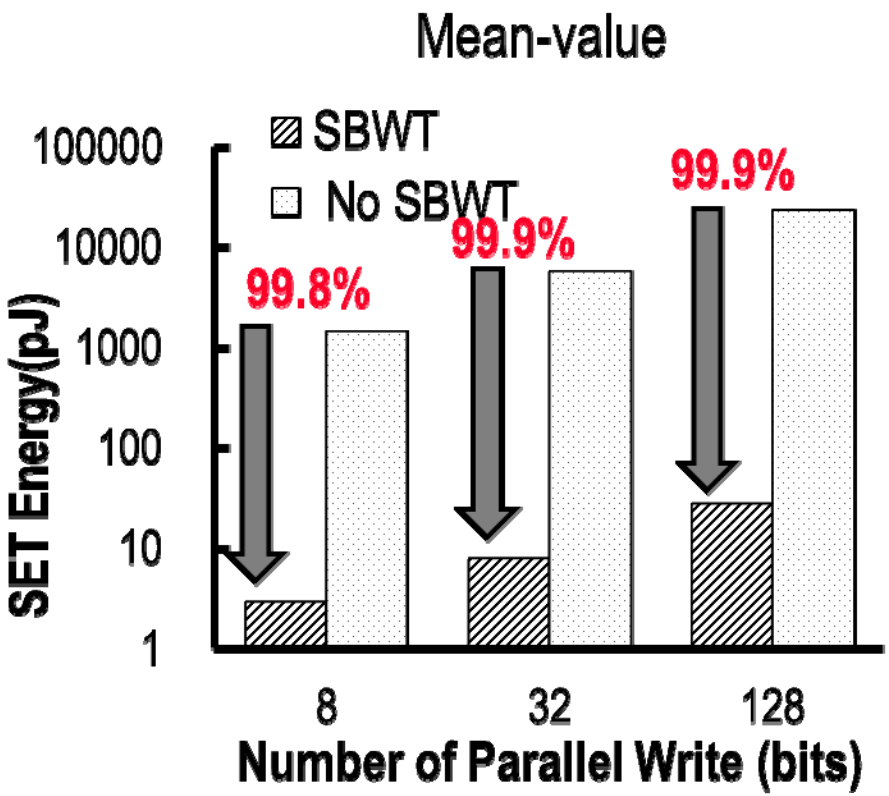




- When HRS→LRS
  - $I_{DC-SET} >> I_{NSD\_MAX} (= k \times I_{SET})$
  - Self-boost V_{DL} to ($V_{WL-SET} - V_{TH-CS}$)
    → cell-switch(CS) is off
    → I_{DC-SET}=0

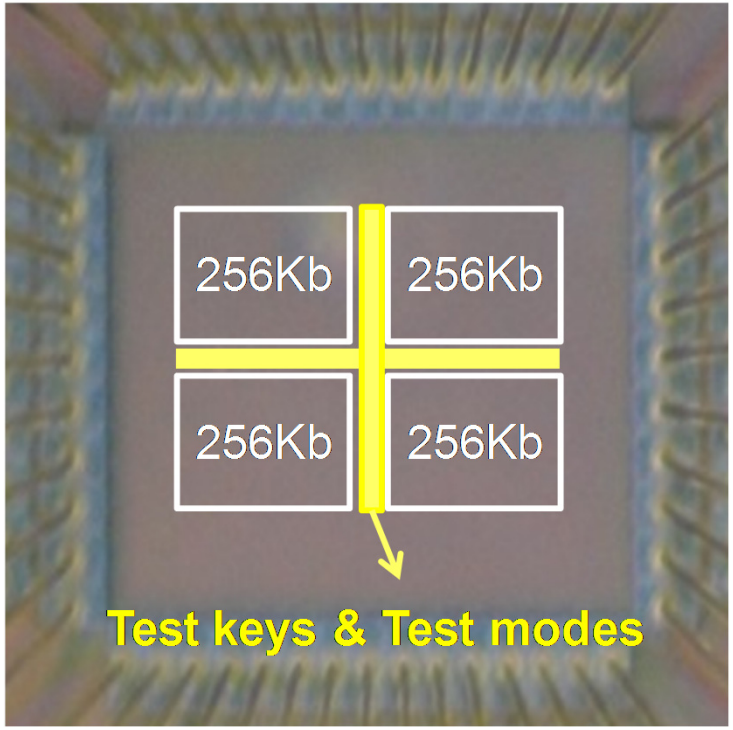
Performance of SBWT





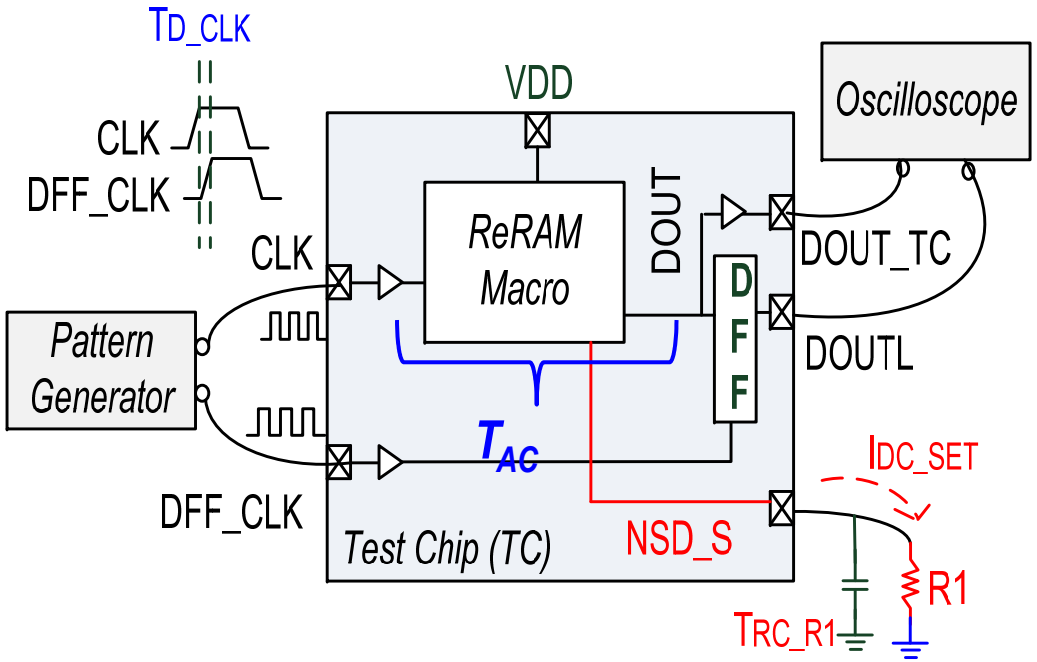
- 99% SET energy reduction for single-bit (10K points)
- 99.9% SET energy reduction for a 32b-IO ReRAM
  - More reduction for wider-IO

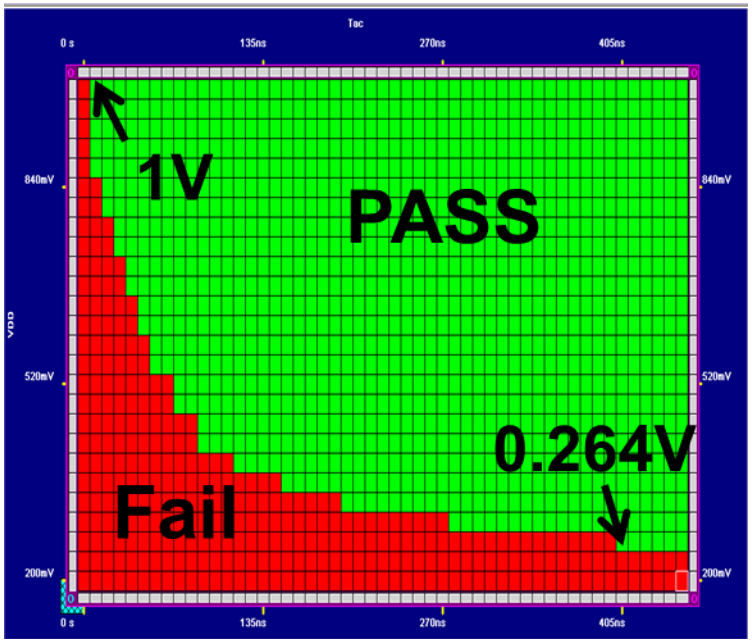
Testchip Summary and Photo

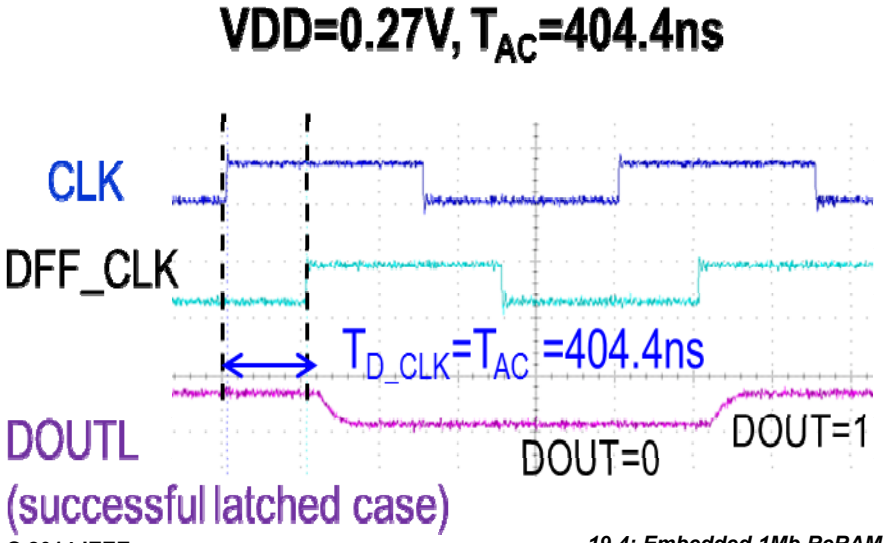


Technology	28nm Hi-K MG CMOS Process
Capacity	1Mb (4 x 256Kb)
Sub-Array (256Kb)	512 rows x 512 columns
ReRAM Cell Area	0.0308um^2
Sun-block Area (256Kb)	0.56mm^2 (include test-modes)
Read Speed	6.8ns at typical VDD (0.85V) 404ns at VDD=0.27V (at LRS>75K ohm) VDD range: 0.27V-1V
Write Speed	SET: ~500ns (at 25uA) RESEST: ~100us (at 50uA)
Read Power	VDD=0.85V: 77uA at 10Mhz VDD=0.27V: 3.2uA at 1Mhz

Measurement Results – Read

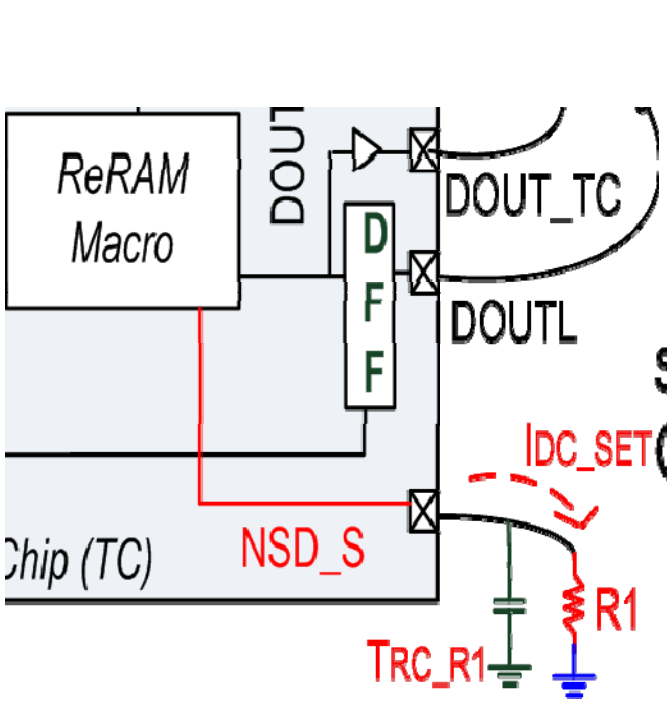


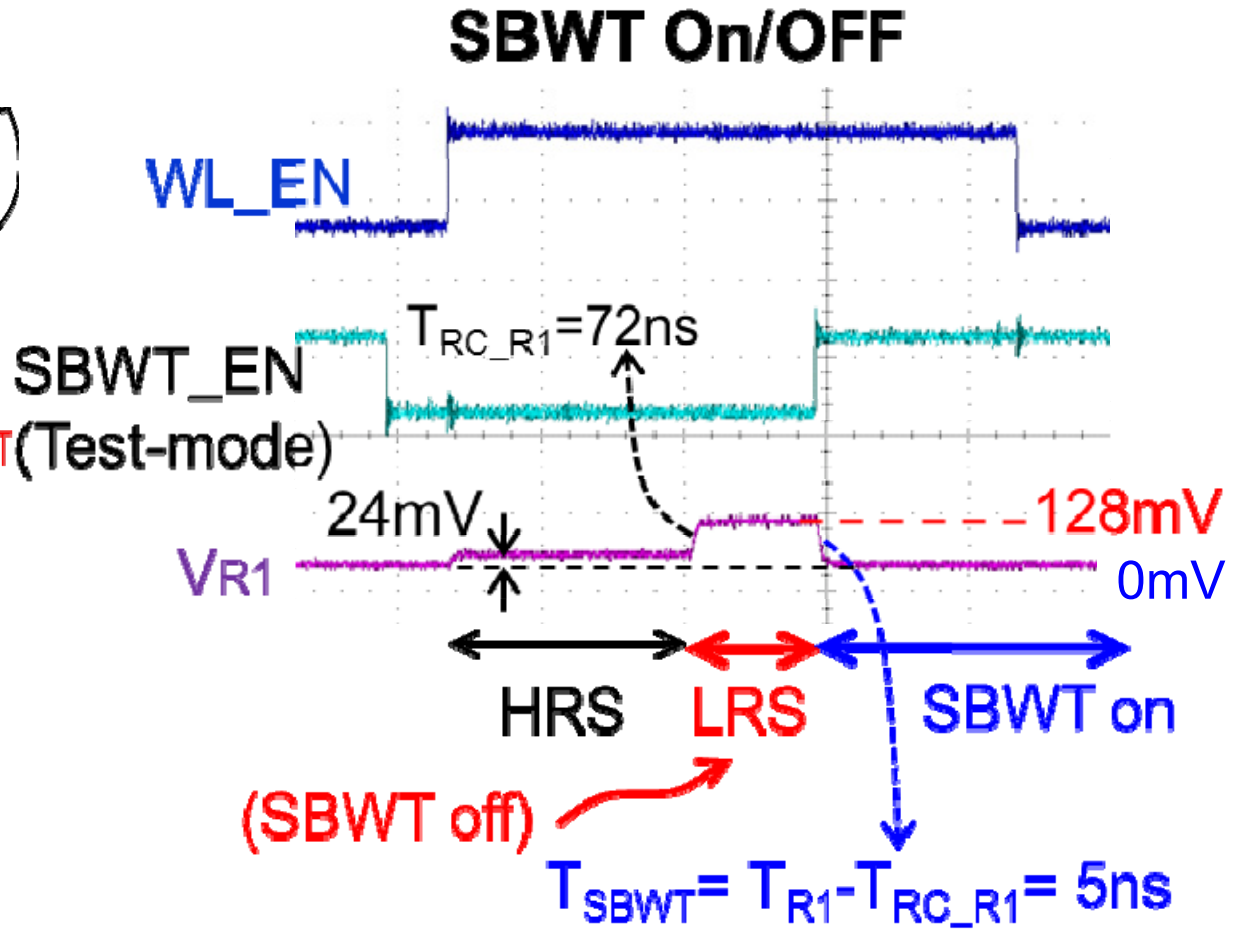




- Use DFF for access time (T_{AC}) extraction
- Read-VDD=1V~0.264V
  - T_{AC}= 6.8ns at 0.85V
  - T_{AC}= 404.4ns at 0.27V

Measured Results - SBWT





- SBWT off: V_{R1} =128mV (8 bits)
- SBWT on: V_{R1} = 0mV (No I_{DC_SET})

Summary

- Swing-Sample-and-Couple Sense Amplifier
  - 1.8+x greater SM for lower read-VDDmin
  - 1.7+x faster T_{AC} across various VDD
- Self-boost-write termination scheme (SBWT)
  - Cut off $I_{DC\text{-}SET}$ of faster-T_{SET} devices
  - Small area penalty < 0.5% (4T)
  - 99+% SET energy reduction
- Verified in a 28nm 1Mb ReRAM macro
  - A logic-process compatible macro
  - Wide VDD range (1V~0.27V)
  - Fast macro read speed
    - ~6.8ns at VDD=0.85V
    - 404.4ns at VDD=0.27V

Thank You for Your Attention

Acknowledgements

Tapeout Service: Flash team of DTP, TSMC

Founding: TSMC-JDP and NSC-Taiwan

Three-Dimensional 128Gb MLC Vertical NAND Flash Memory with 24-WL Stacked Layers and 50MB/s High-Speed Programming

Ki-Tae Park, Jin-man Han, Daehan Kim, Sangwan Nam, Kihwan Choi, Min-Su Kim, Pansuk Kwak, Doosub Lee, Yoon-He Choi, Kyung-Min Kang, Myung-Hoon Choi, Dong-Hun Kwak, Hyun-wook Park, Sang-won Shim, Hyun-Jun Yoon, Doohyun Kim, Sang-won Park, Kangbin Lee, Kuihan Ko, Dong-Kyo Shim, Yang-Lo Ahn, Jeunghwan Park, Jinho Ryu, Donghyun Kim, Kyungwa Yun, Joonsoo Kwon, Seunghoon Shin, Dongkyu Youn, Won-Tae Kim, Taehyun Kim, Sung-Jun Kim, Sungwhan Seo, Hyung-Gon Kim, Dae-Seok Byeon, Hyang-Ja Yang, Moosung Kim, Myong-Seok Kim, Jinseon Yeon, Jaehoon Jang, Han-Soo Kim, Woonkyung Lee, Duheon Song, Sungsoo Lee, Kye-Hyun Kyung, Jeong-Hyuk Choi, Kinam Kim

Samsung, Hwasung, Korea

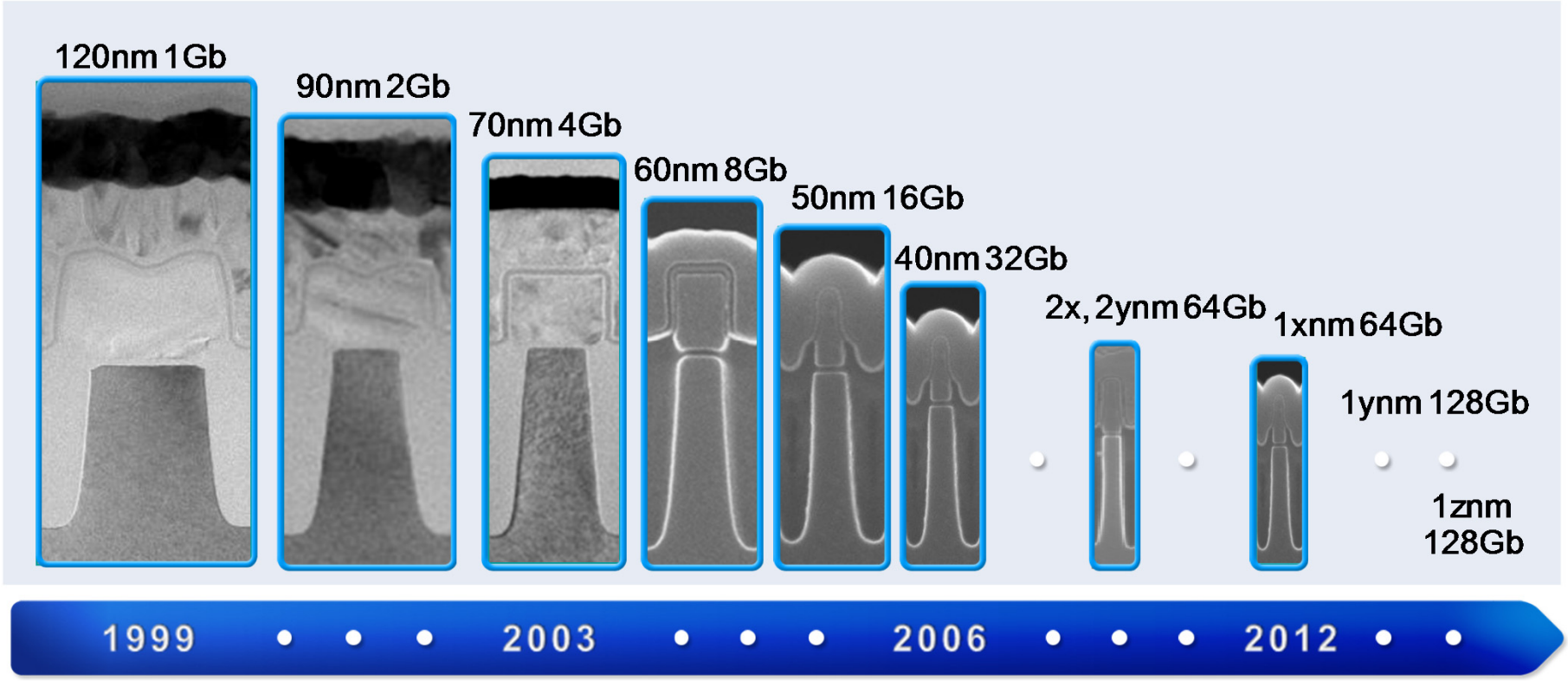
Outline

- **Introduction**
  : Planar NAND vs. V-NAND
- **Technology & Chip Architecture**
- **Designs for V-NAND**
- **Performance & Power**
- **Summary**

Planar NAND Scaling

✓ Planar NAND Evolution over 20 years

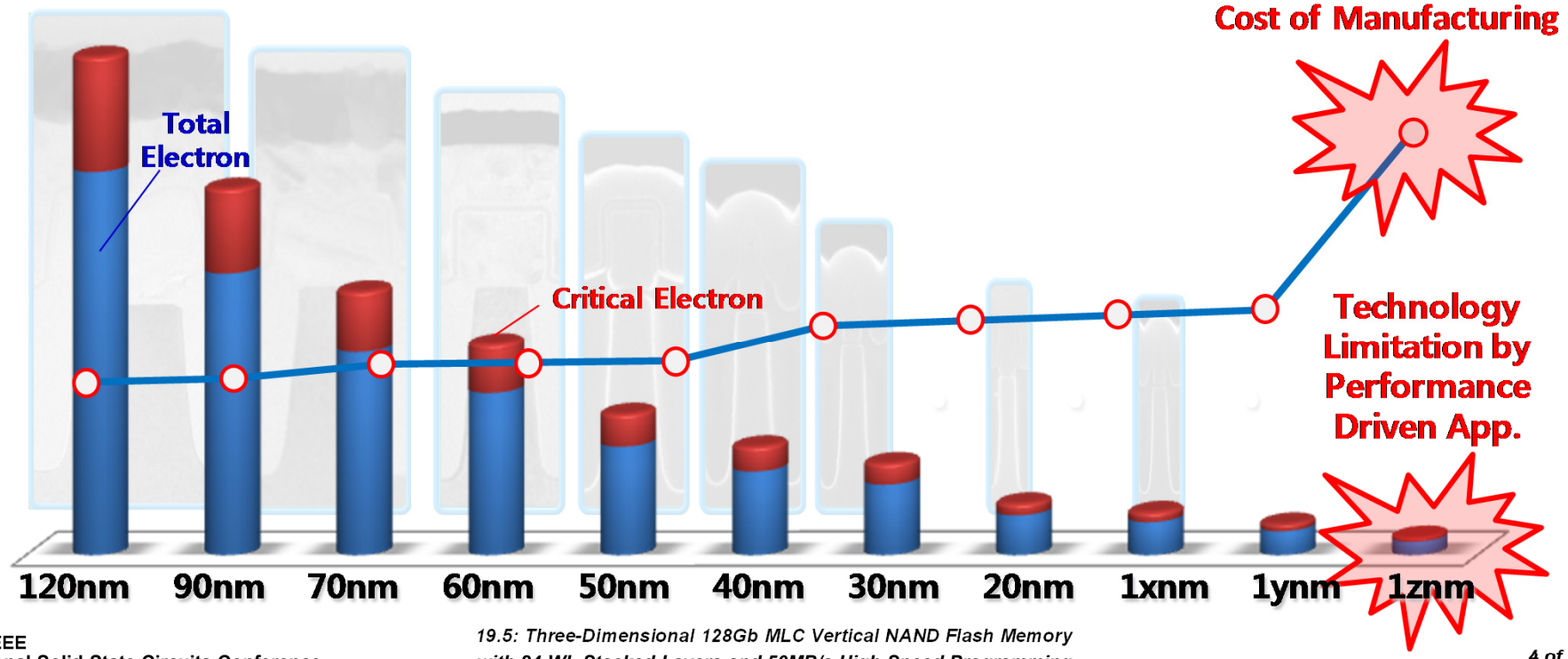
✓ Now, NAND Scaling under 1ynm is facing Critical Challenges



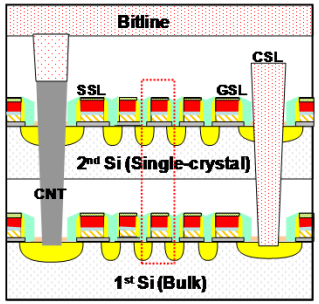
Uncertainty of Future Planar NAND

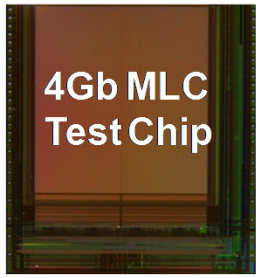
✓ Quadruple Patterning + Sophisticated Processes (WL/Active Air-gap, Metal-gate etc.) at 1znm-node
→ Drastically Increased Manufacturing Cost

✓ Few Storage Electrons + Large Cell-to-cell Interference
→ Technology Limitation by High Performance-driven NAND Applications (Mobile, SSD etc.)



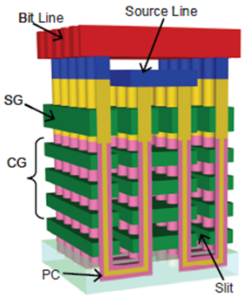
Developing 3-Dimensional NAND

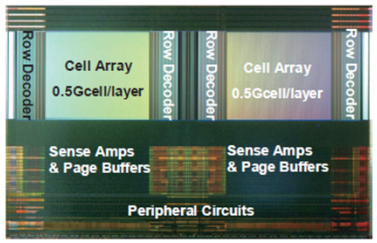




3D Stacked NAND
(IEDM'06, ISSCC'08)

3D Vertical NAND, BiCS
(VLSI'07, VLSI'09)

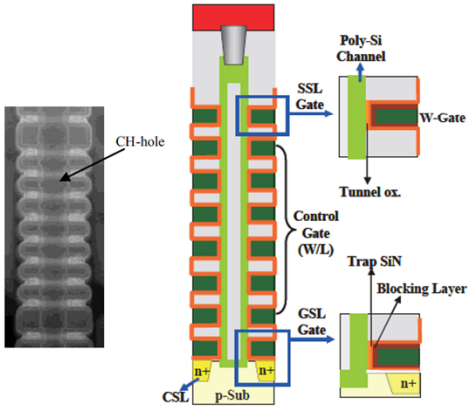




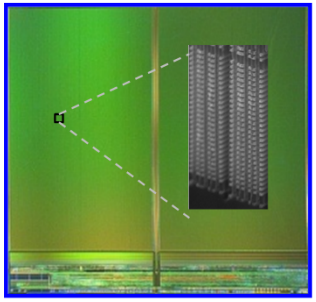
From bottom to top over 10 years

1. Material Innovation
2. Structure Innovation
3. Integration Innovation
4. Design Innovation
5. Managing Innovation

3D Vertical NAND, TCAT
(VLSI'09)

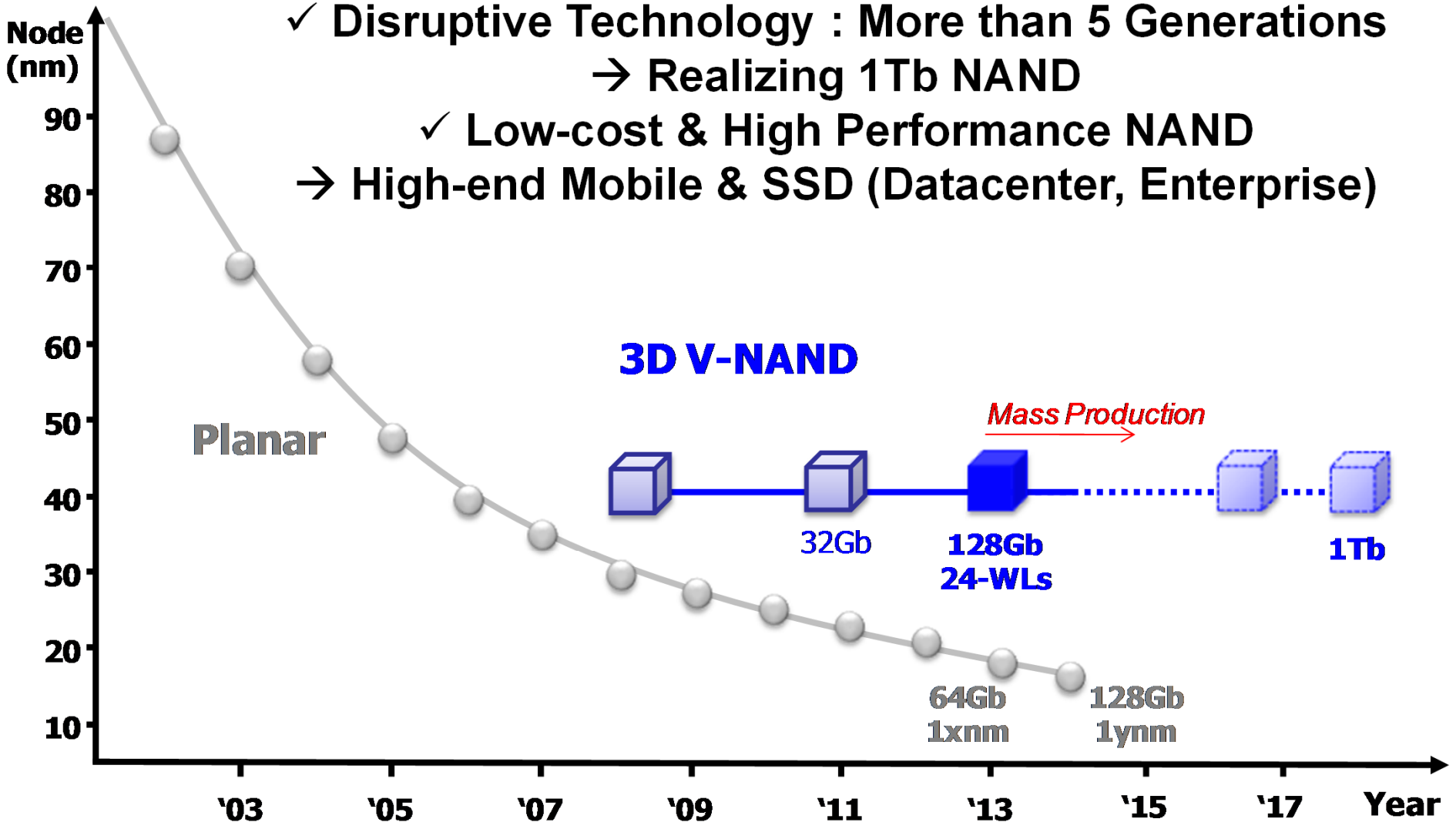


128Gb MLC
3D Vertical NAND
Product,
(This Work)



Next NAND Flash Era

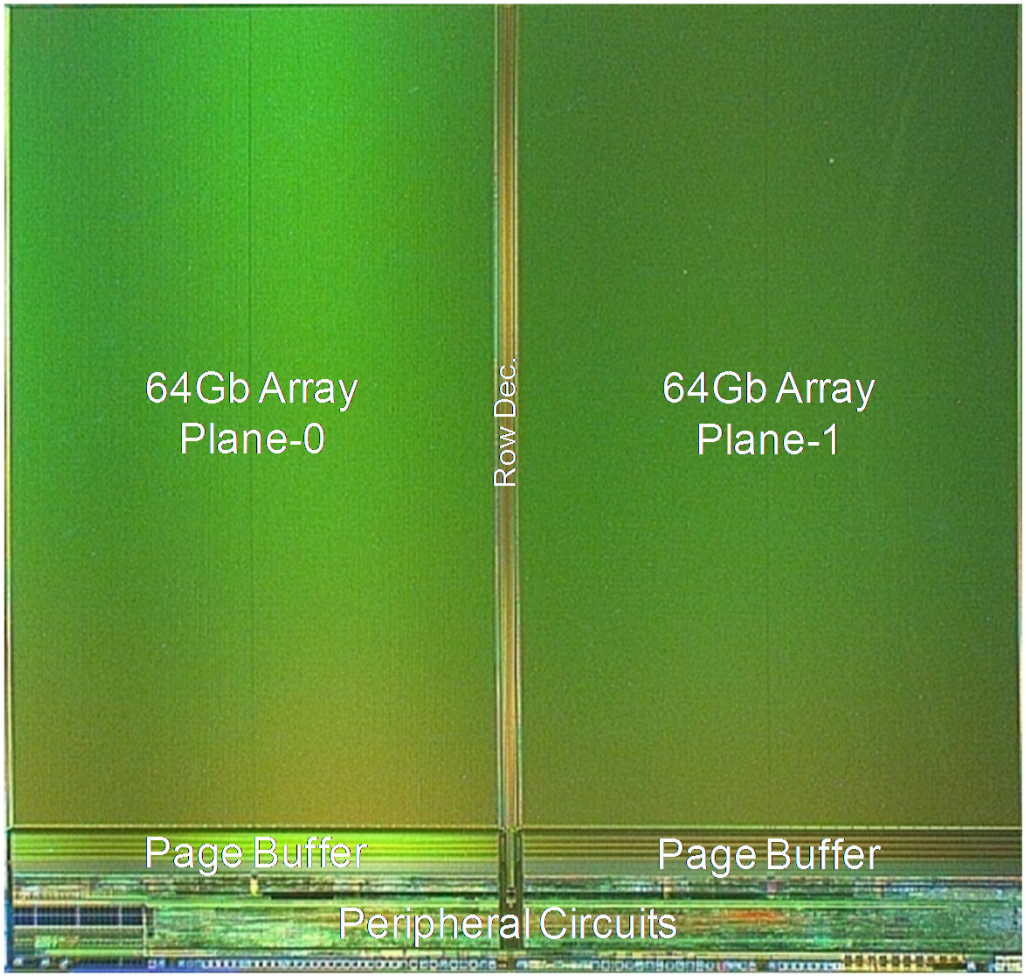
✓ Disruptive Technology : More than 5 Generations
→ Realizing 1Tb NAND
✓ Low-cost & High Performance NAND
→ High-end Mobile & SSD (Datacenter, Enterprise)



Outline

- Introduction
  : Planar NAND vs. V-NAND
- **Technology & Chip Architecture**
- Designs for V-NAND
- Performance & Power
- Summary

128Gb MLC V-NAND Product



- Vertical-NAND Technology

- Chip Size
  : 133mm^2 → 0.96Gb/mm^2

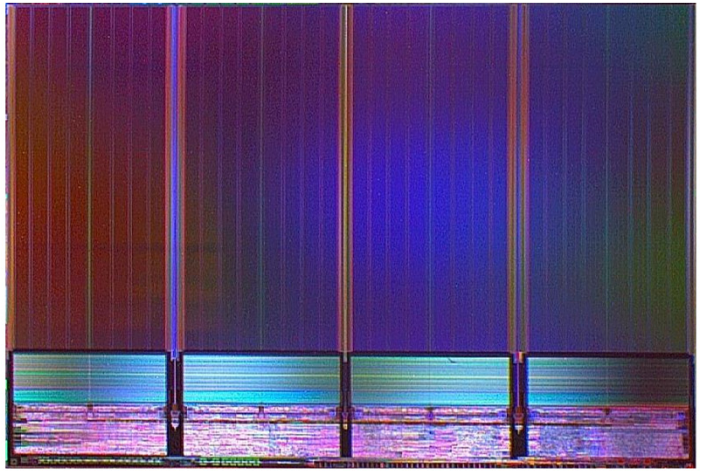
- 24-WL Stacked Layers

- 64Gb Array × 2-Plane

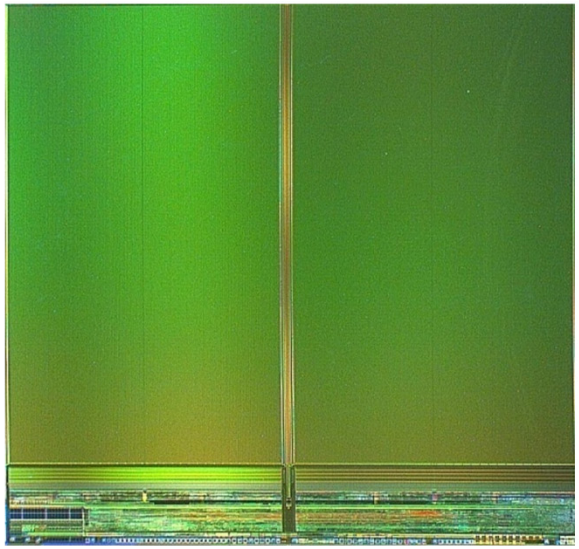
- One-sided Page Buffer
  : (8KB x 2) Page Size

- Asynchronous DDR Interface
  : Wave-pipeline datapath
  : 667Mbps at Mono Die
  : 533Mbps at 8-stacked Dies

Device Comparison



×2 Density Increasing



1xnm Planar MLC (ISSCC'12)	vs.	24-WLs V-NAND (This Work)
64Gb	Memory Size	128Gb
32KB One-sided	Page Buffer	16KB One-sided
33MB/s	Performance	50MB/s
Wave-pipeline, 533Mbps	IO Speed	Wave-pipeline, 667Mbps
109.5mm^2	Chip Size	133mm^2
0.585Gb/mm^2	Density	0.96Gb/mm^2

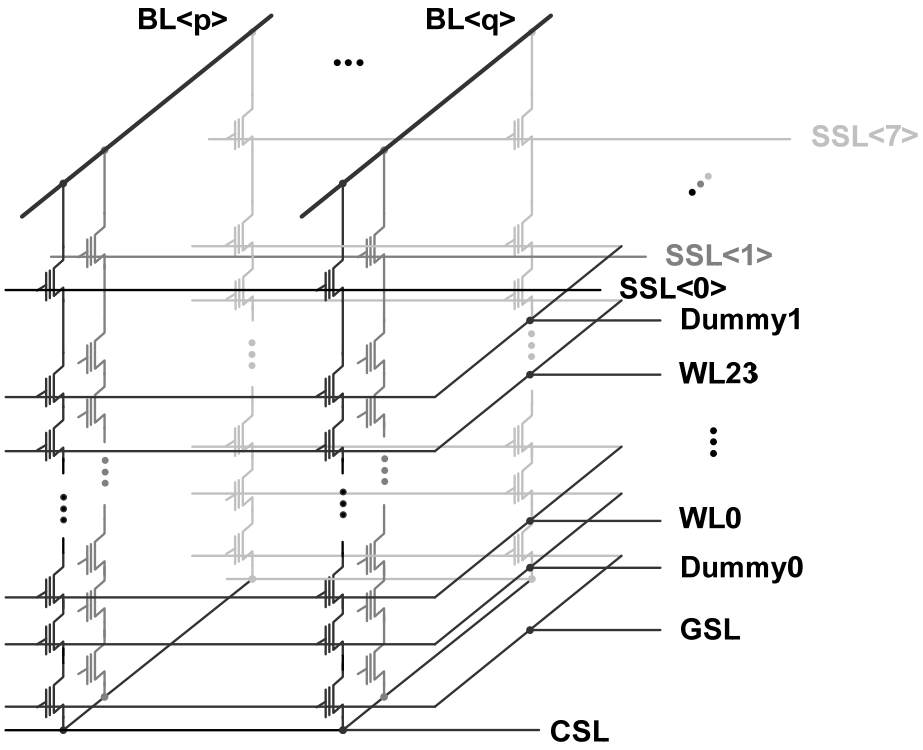
x1.5↑

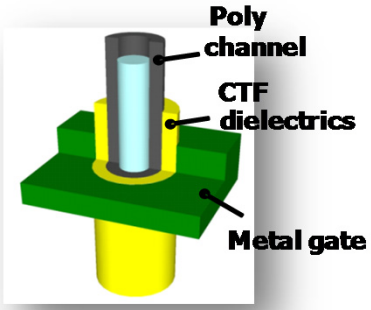
x1.25↑

x1.64↑

V-NAND Array Structure

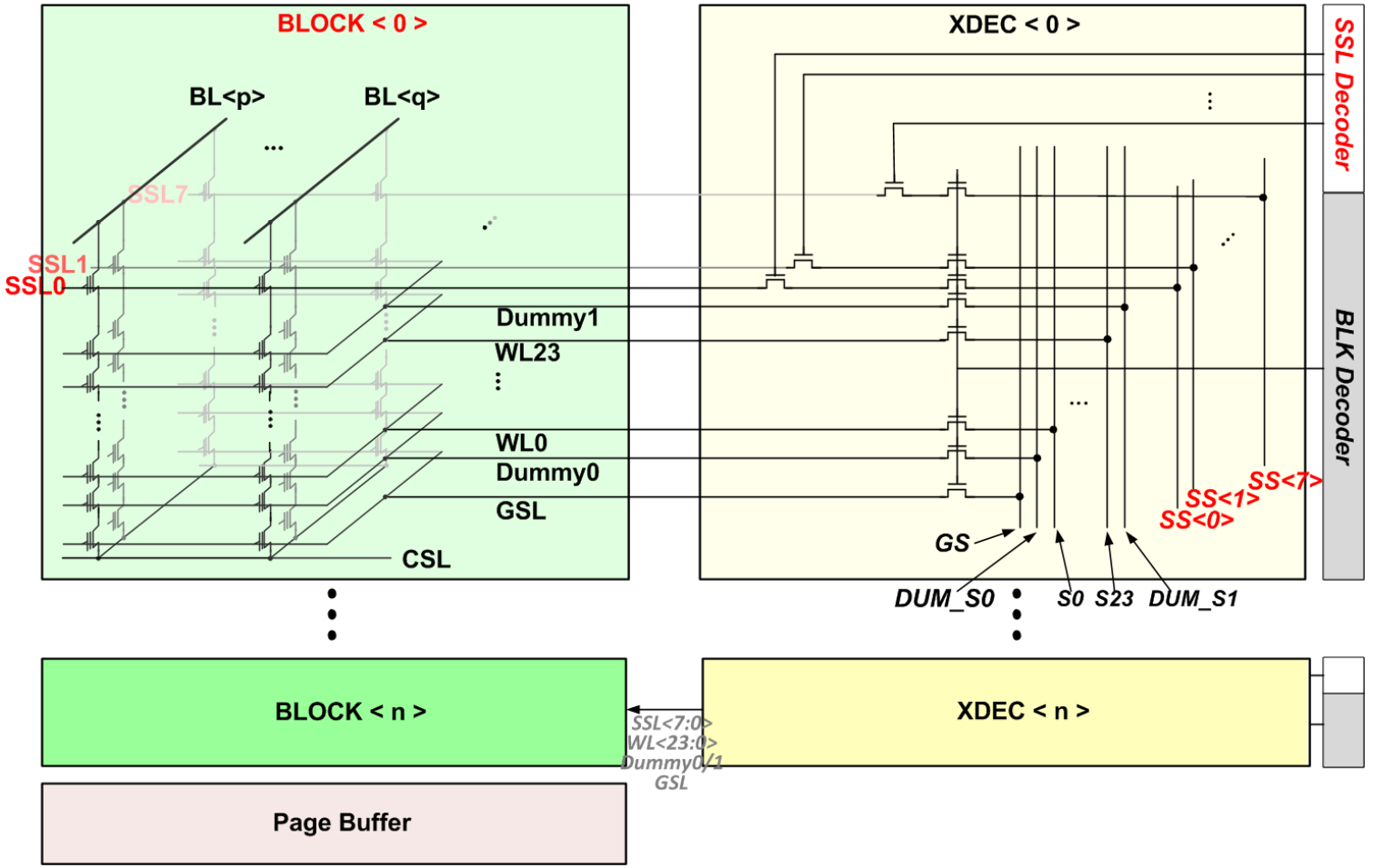
- ✓ Advanced V-NAND Technology with Damascened Metal Gate
- ✓ Cell : All-around Gate Structure + Charge Trap Flash
- ✓ String : 24-WL + 2-DWL + 2-Select WL
- ✓ Block : 8 Strings with Shared BL (8KB)





V-NAND Core Architecture

✓ **Page Program & Bulk Erase using FN Tunneling like 2D NAND**
→ Modified Conventional 2D NAND Core Architecture

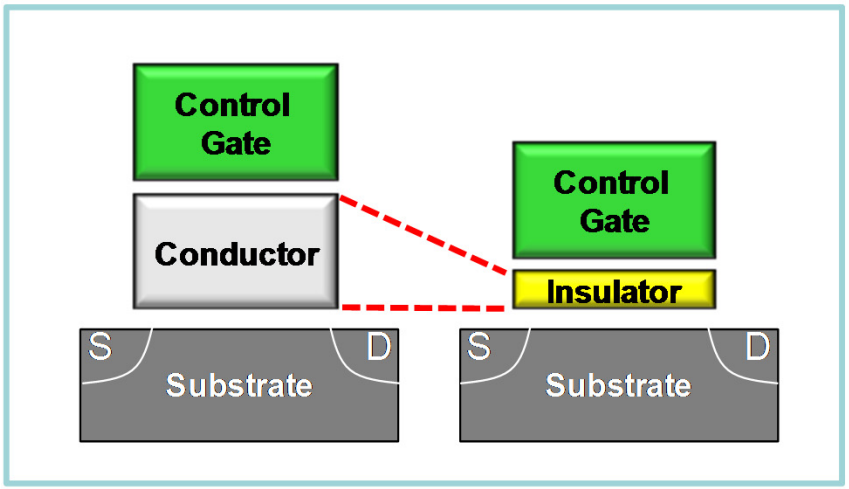


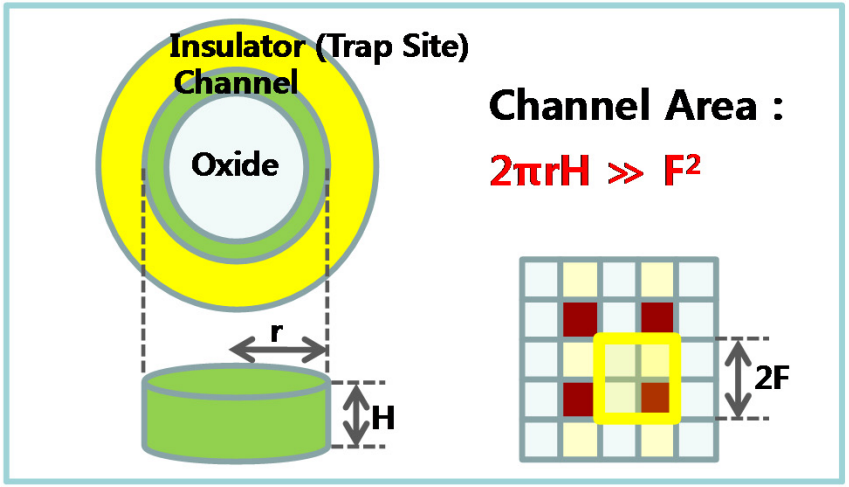
Outline

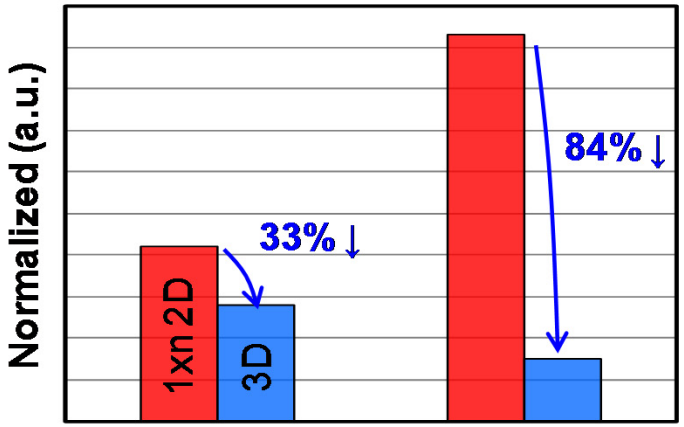
- Introduction
  : Planar NAND vs. V-NAND
- Technology & Chip Architecture
- **Designs for V-NAND**
- Performance & Power
- Summary

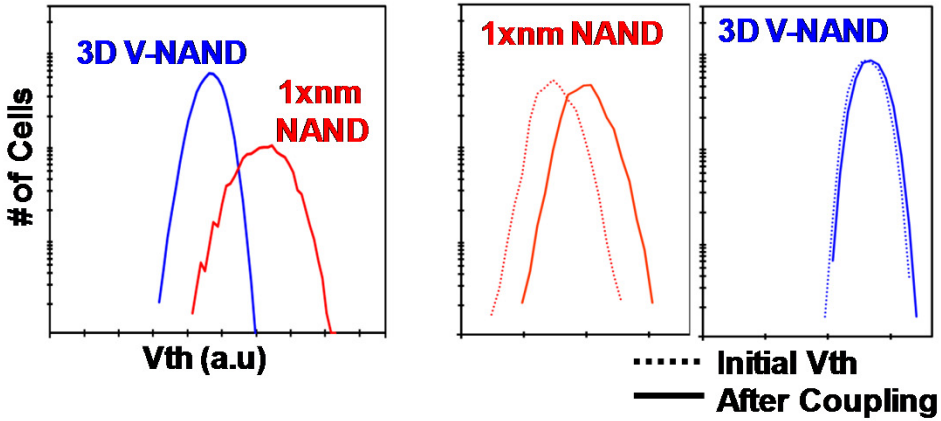
Cell Characteristic Comparison

✓ Advanced CTF + All-around Gate Structure
→ Superior Cell Characteristics



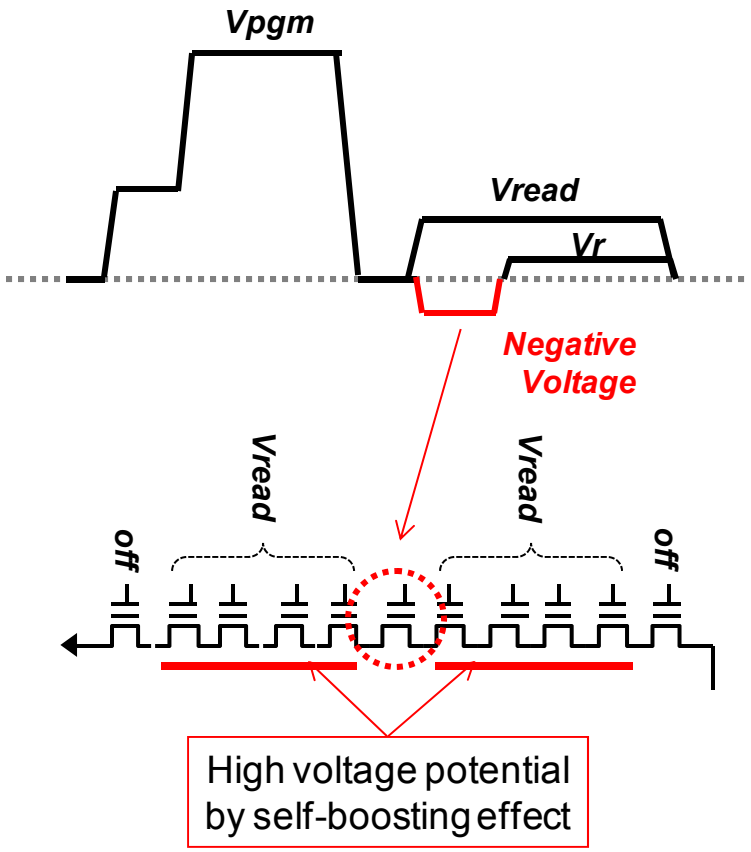


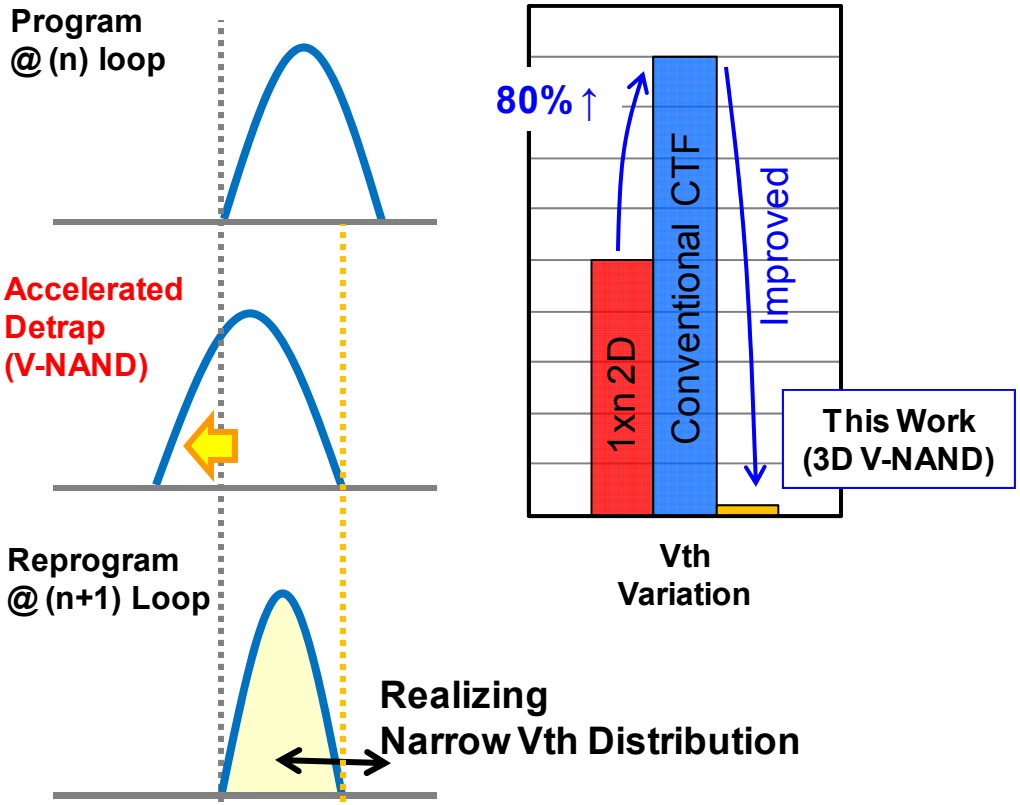




Counter-pulse using Self-boosting

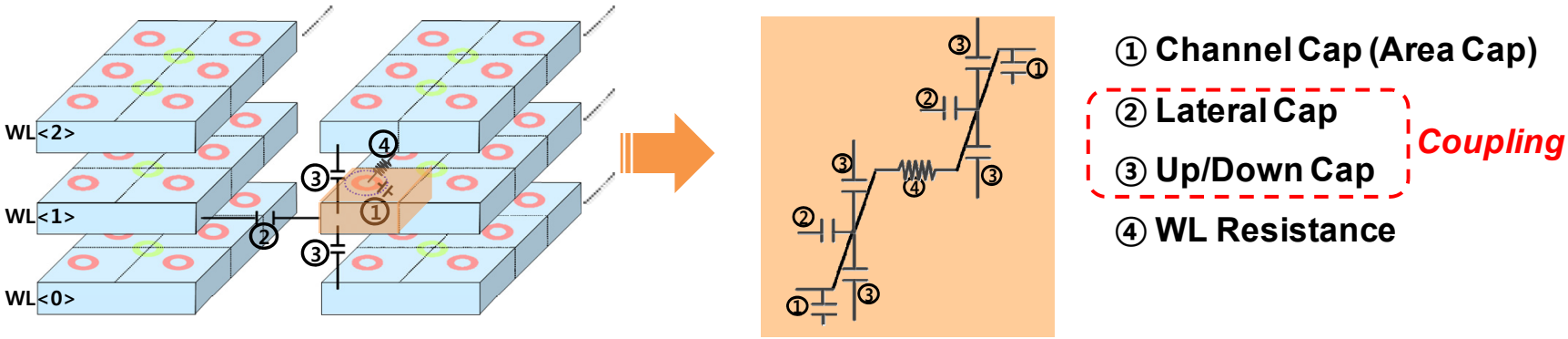
✓ **Electrical Accelerating Detrap + Optimized Device Engineering**
→ Better Vth Distribution can be realized

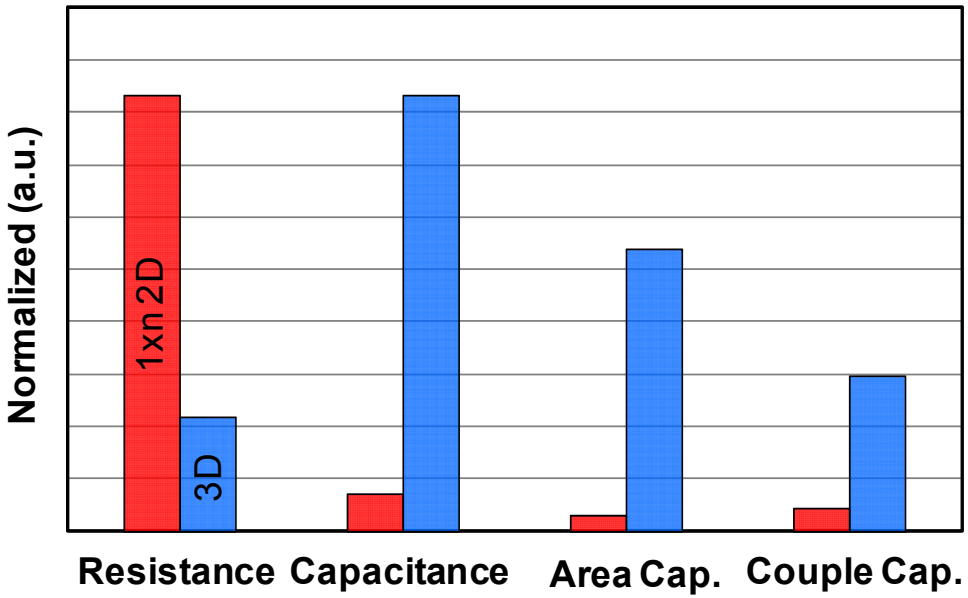




Signal Crosstalk of V-NAND Array

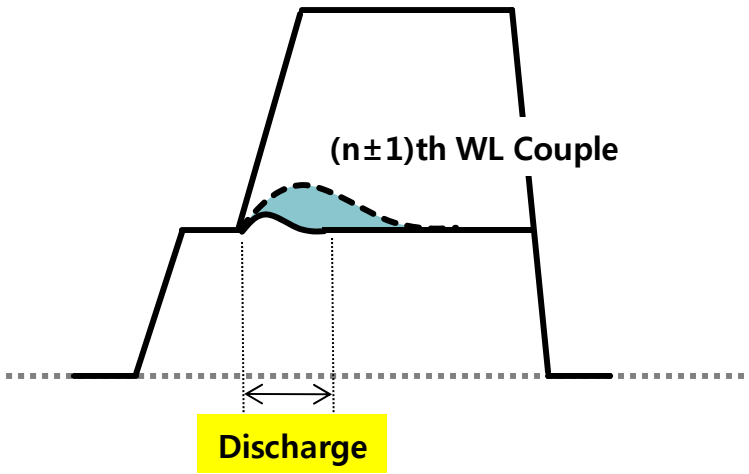
✓ Large Capacitive Coupling between WLs
→ Design for Signal Integrity of 3D Memory Array



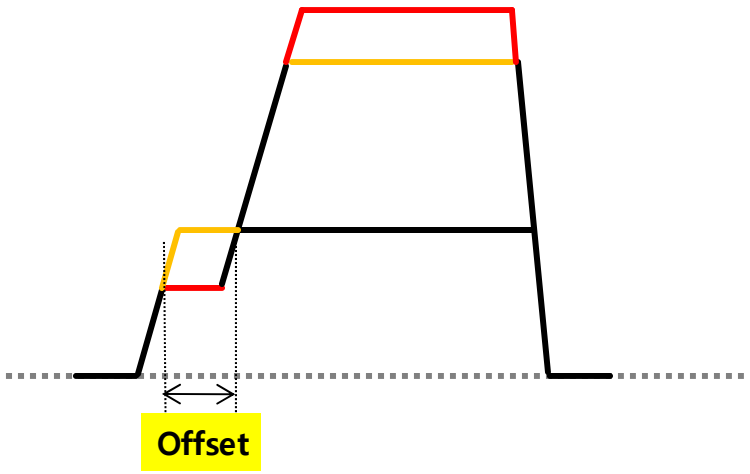


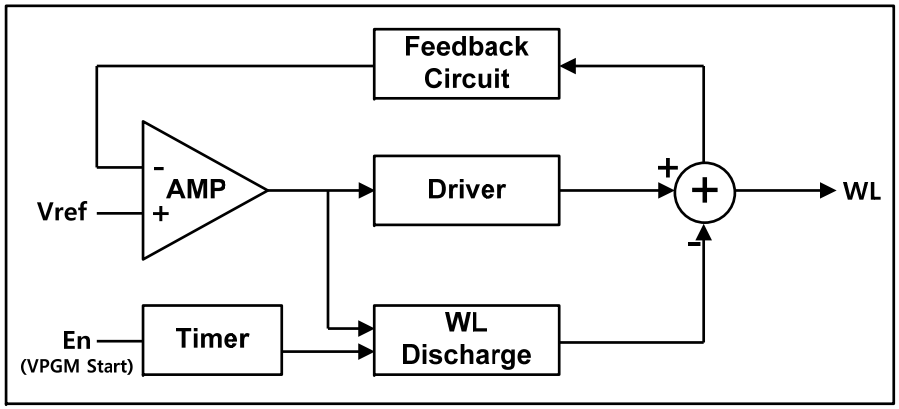
Designs for Core Signal Integrity

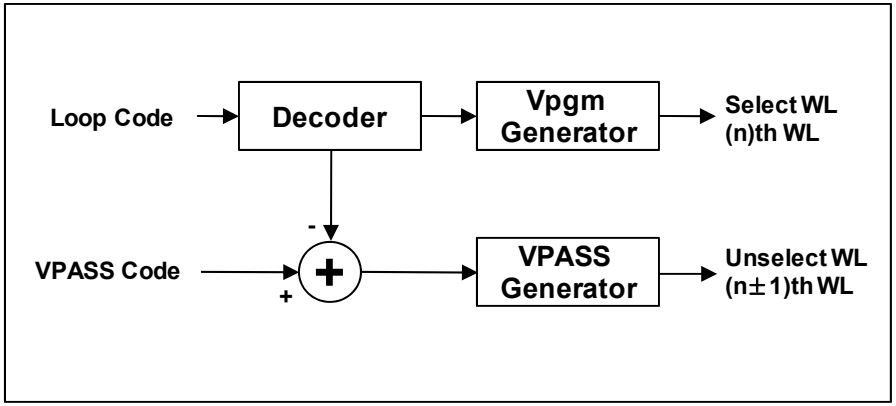
WL Degeneration Scheme



Coupling-predicted Offset Control Scheme

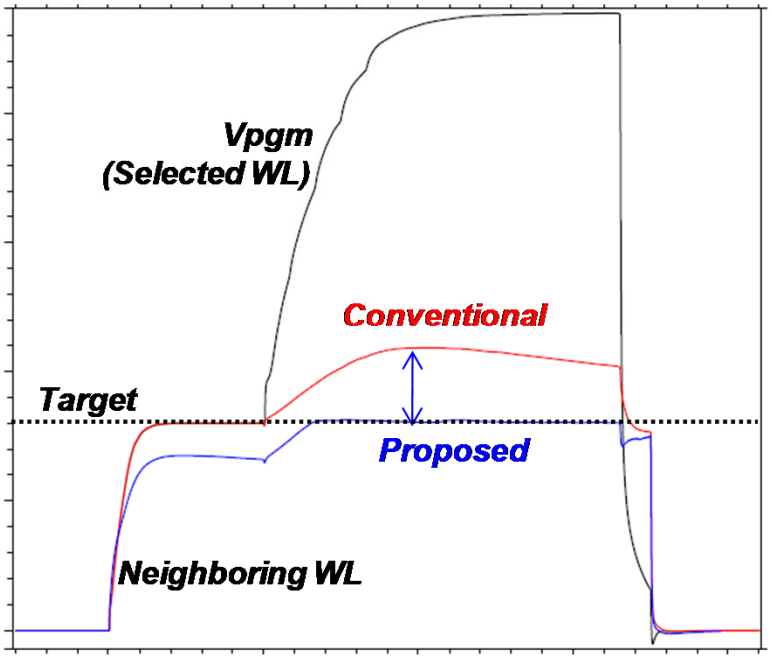


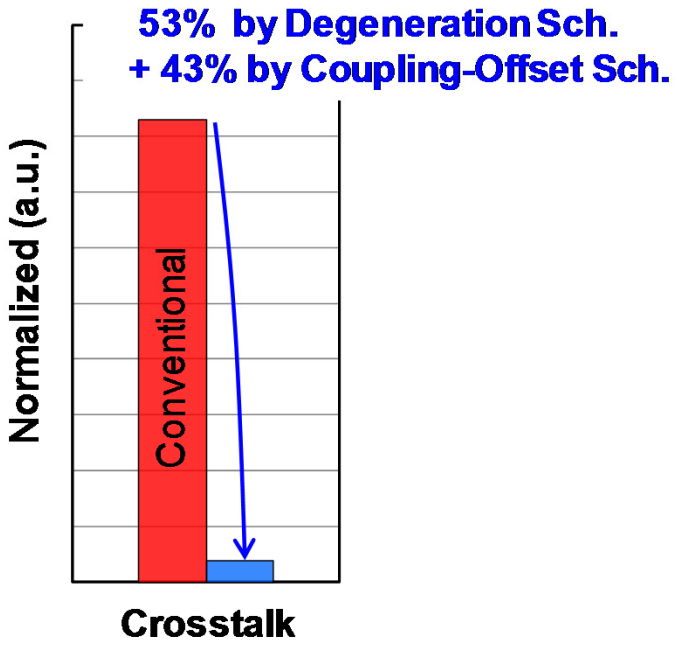




Results of Reduction Schemes

✓ Well-controlled Overshoot of Victim WL
→ Abnormal Disturbance by WL-WL Cross-talk can be dramatically reduced





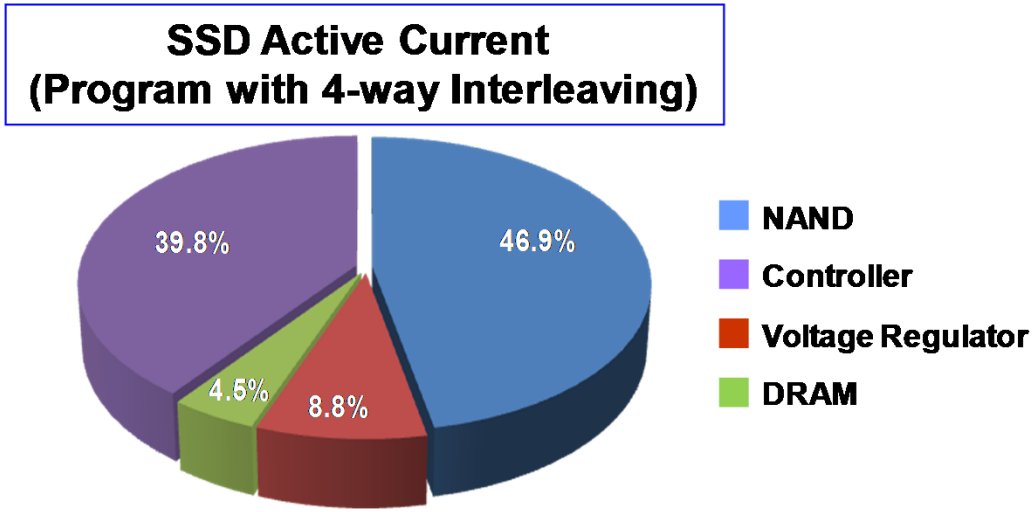
Outline

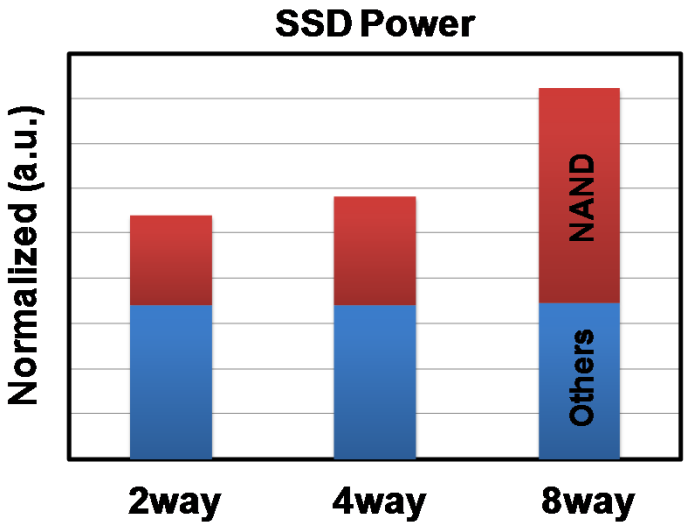
- Introduction
  : Planar NAND vs. V-NAND
- Technology & Chip Architecture
- Designs for V-NAND
- **Performance & Power**
- Summary

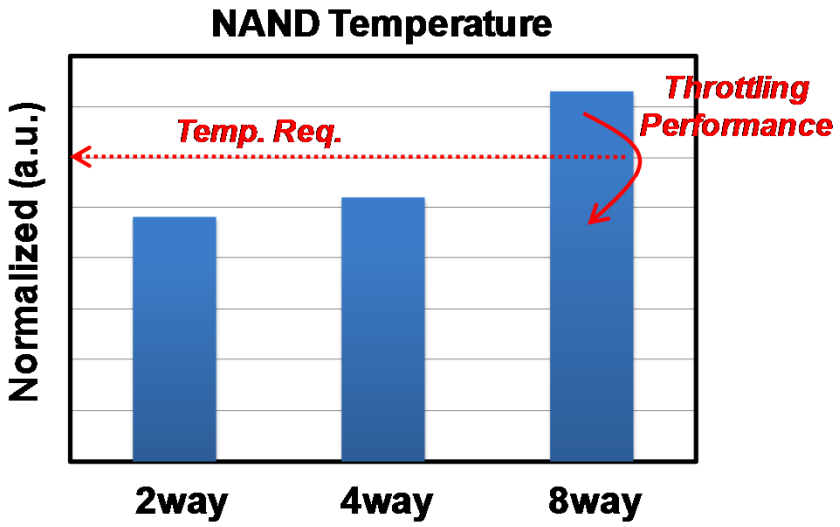
NAND System Power

✓ SSD Power → SSD Temp. → Throttling Performance

✓ Lowering NAND Power → Increasing SSD Performance

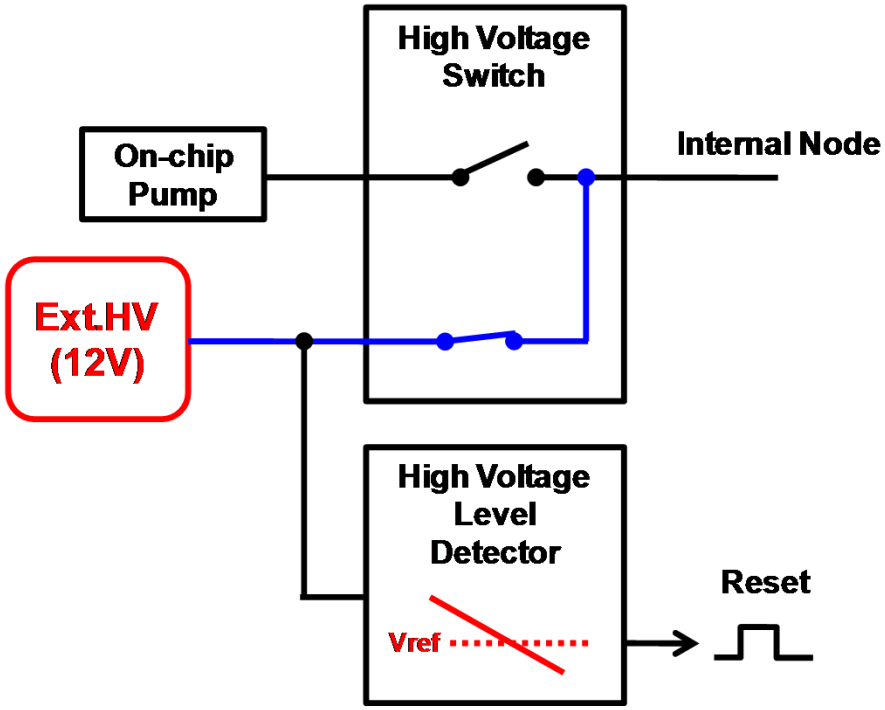


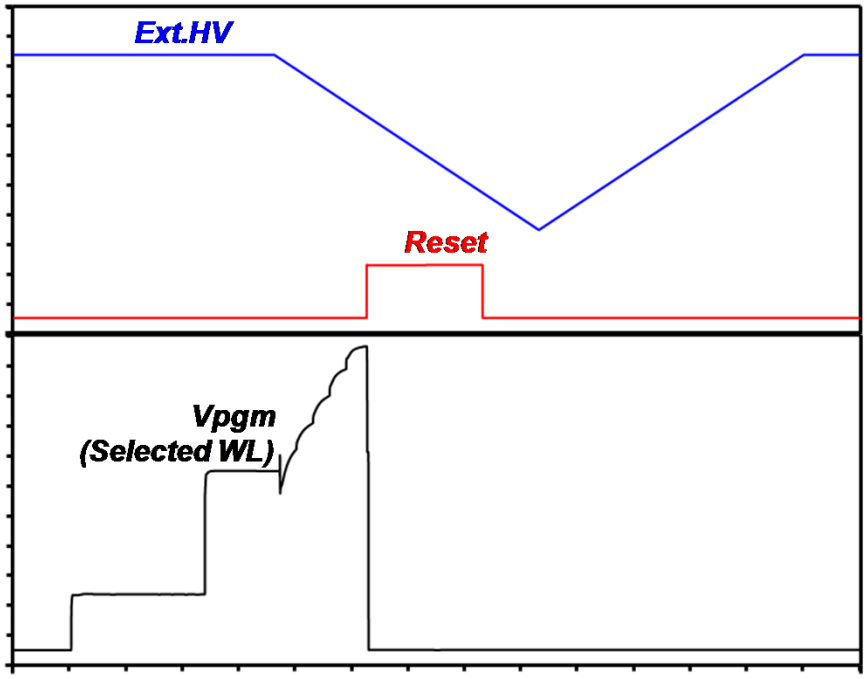




External High-voltage Supply Scheme

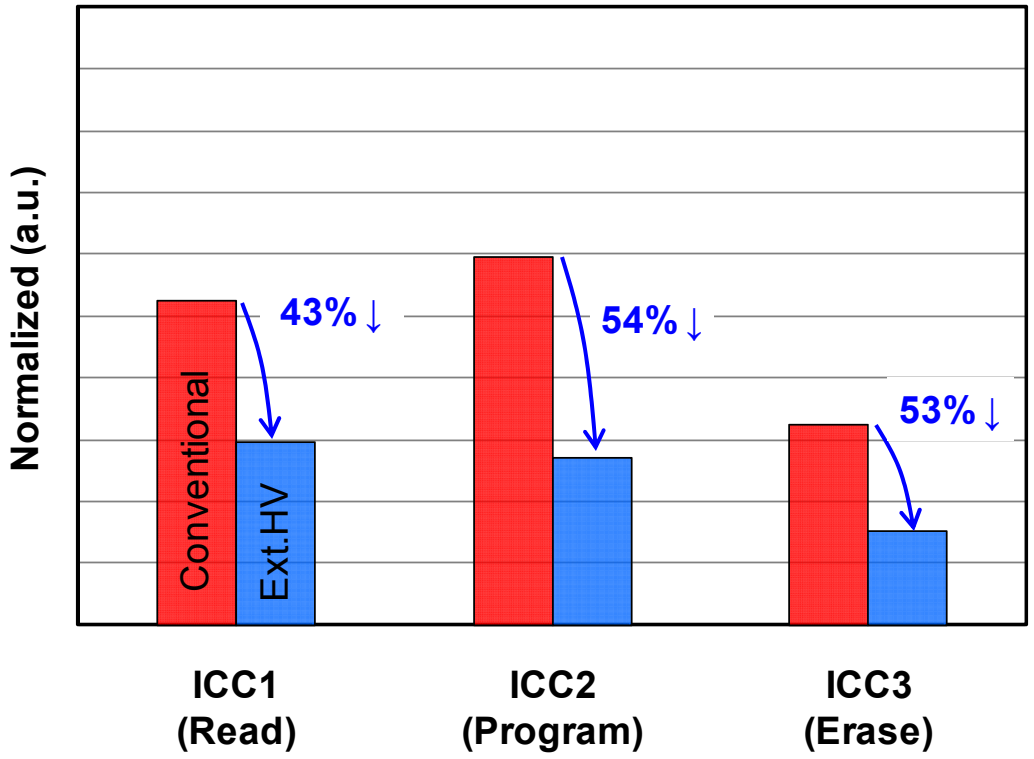
- ✓ External High-voltage (eg.12V) can be used as on-chip Pump's Source for more efficiency
- ✓ Protecting Internal Circuit Scheme by monitoring Ext.HV





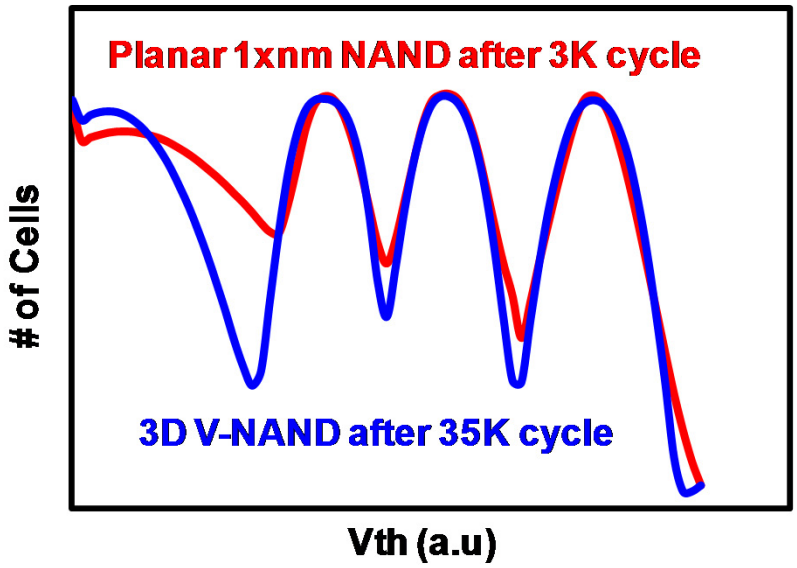
Measured Active Power

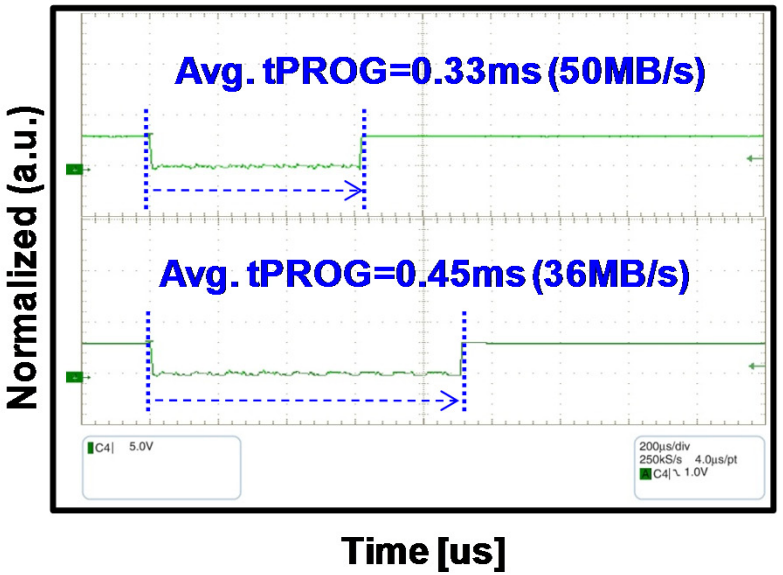
✓ Over 50% Lower Energy Advantage is achieved
→ Increasing overall SSD Performance
by using 8-way Interleaving NAND Operation



Measured MLC Vth Distribution

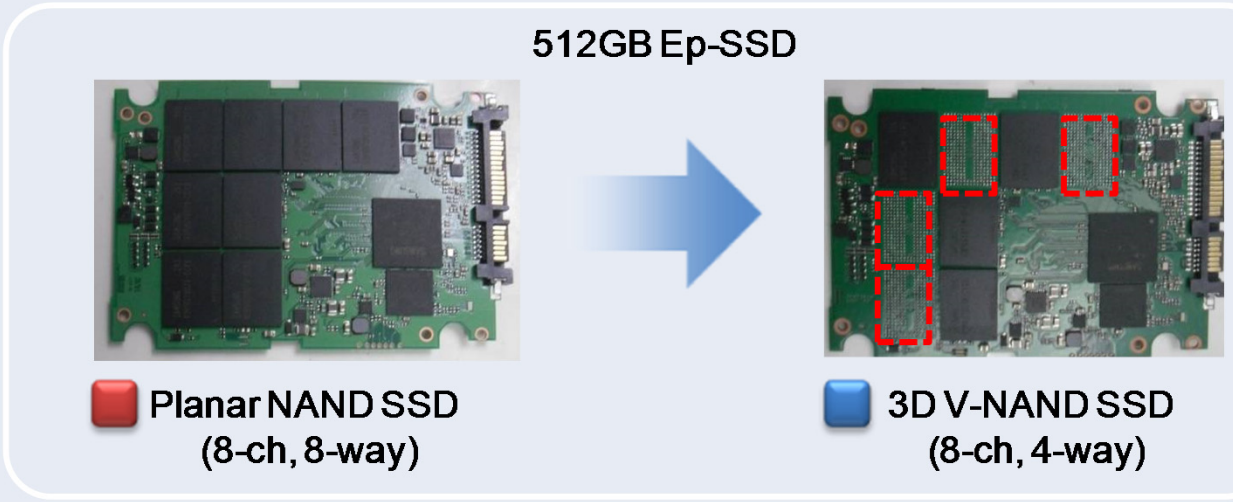
✓ 36MB/s + 35K Endurance
for Data-center & Enterprise SSD Applications
✓ 50MB/s + 3K Endurance for Mobile Applications



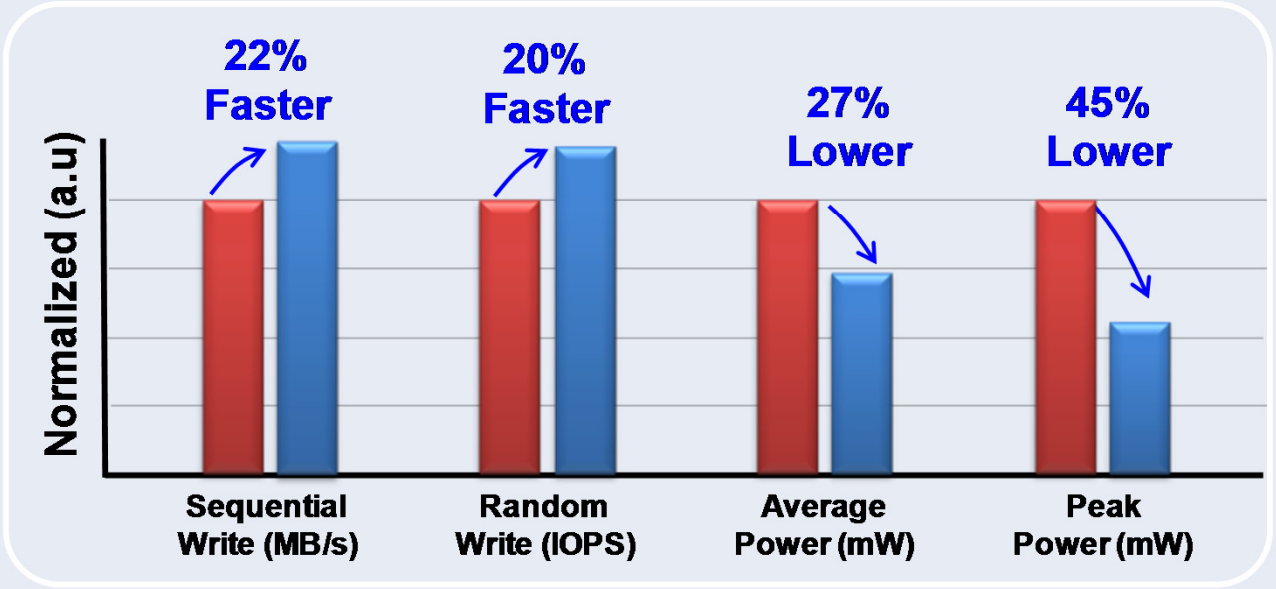


Enterprise SSD Comparison

Smaller Real Estate



Higher Performance



Outline

- Introduction
  : Planar NAND vs. V-NAND
- Technology & Chip Architecture
- Designs for V-NAND
- Performance & Power
- **Summary**

Features

Bits per Cell	2
Density	128Gb
Technology	Three Dimensional Vertical NAND, 3-metals
Organization	8KB × 384 pages × 5464 blocks × 8
Program Performance	50MB/s for Embedded App., 36MB/s for Enterprise SSD
Data Interface Speed	667Mbps@Mono, 533Mbps@8-stack
Power Supply	Vcc=3.3V / Vccq=1.8V

Conclusions

- A True Breakthrough Flash Device is introduced
  : 3D 128Gb MLC vertical NAND (V-NAND) flash memory
  : 24-WL stacked layers

- Higher Performance & Reliability & Lower Energy
  : New programming, WL Cross-talk reduction etc.
  : 36MB/s + 35K endurance for high-end SSD
  : 50MB/s + 3K endurance for mobile
  : >50% energy reduction by external pump source

- Mass Production & Next Flash Era
  : solving defect & achieving high yield in the 3D structure
  : continuing stacking more than next 5 generations

Hybrid Storage of ReRAM/TLC NAND Flash with RAID-5/6 for Cloud Data Centers

Shuhei Tanakamaru[1, 2], Hiroki Yamazawa[1], Tsukasa Tokutomi[1],
Sheyang Ning[1,2], and Ken Takeuchi[1]

[1]Chuo University
[2]University of Tokyo

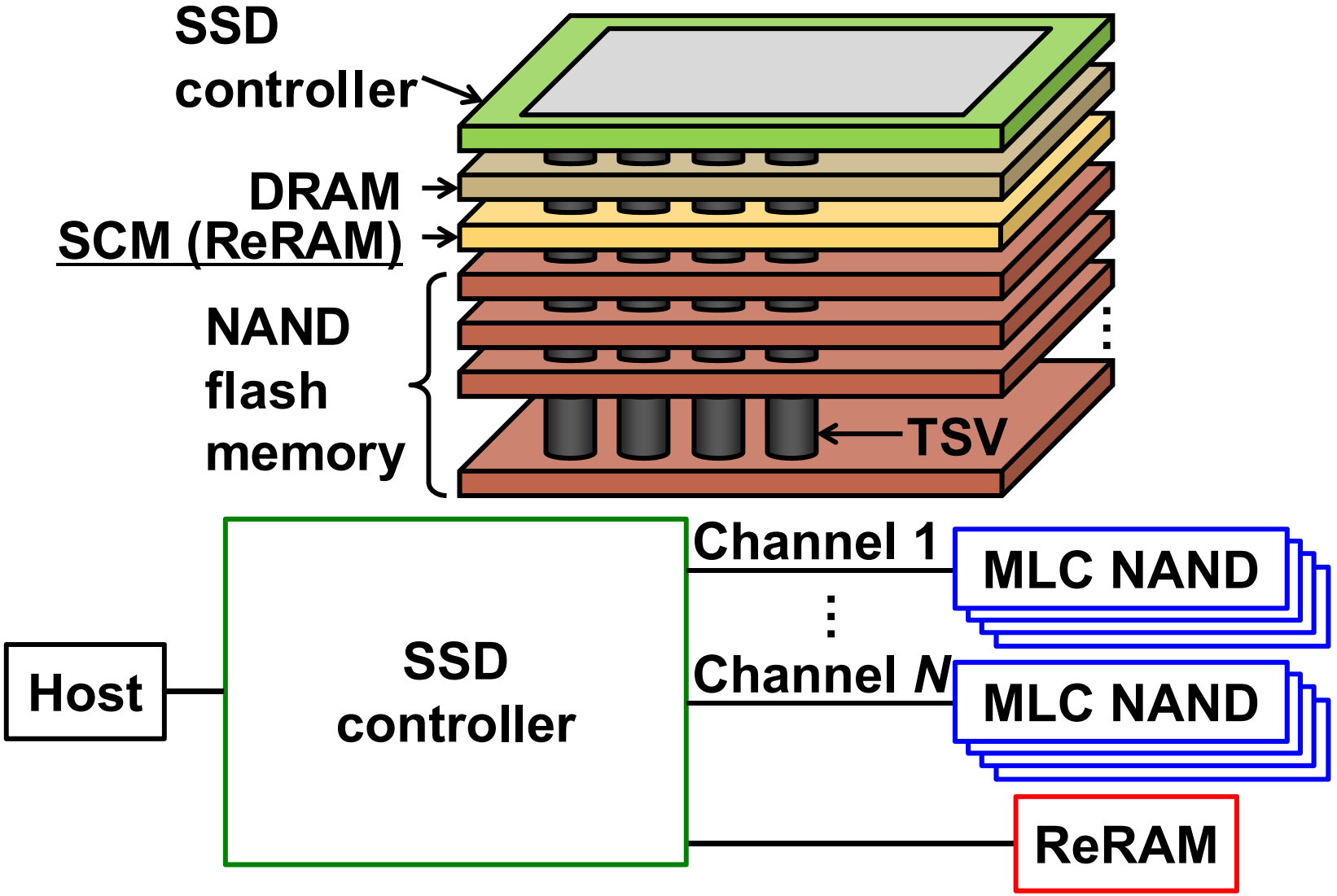
Outline

- **Introduction**
- **Reliability Improvement of ReRAM**
  - **Flexible R_{Ref} (FR)**
  - **Adaptive Asymmetric Coding (AAC)**
  - **Verify Trials Reduction (VTR)**
- **Reliability Improvement of NAND Flash Memories**
  - **Balanced RAID-5/6**
  - **Bits/Cell Optimization (BCO)**
- **Summary**

Outline

- **Introduction**
- **Reliability Improvement of ReRAM**
  - **Flexible R_{Ref} (FR)**
  - **Adaptive Asymmetric Coding (AAC)**
  - **Verify Trials Reduction (VTR)**
- **Reliability Improvement of NAND Flash Memories**
  - **Balanced RAID-5/6**
  - **Bits/Cell Optimization (BCO)**
- **Summary**

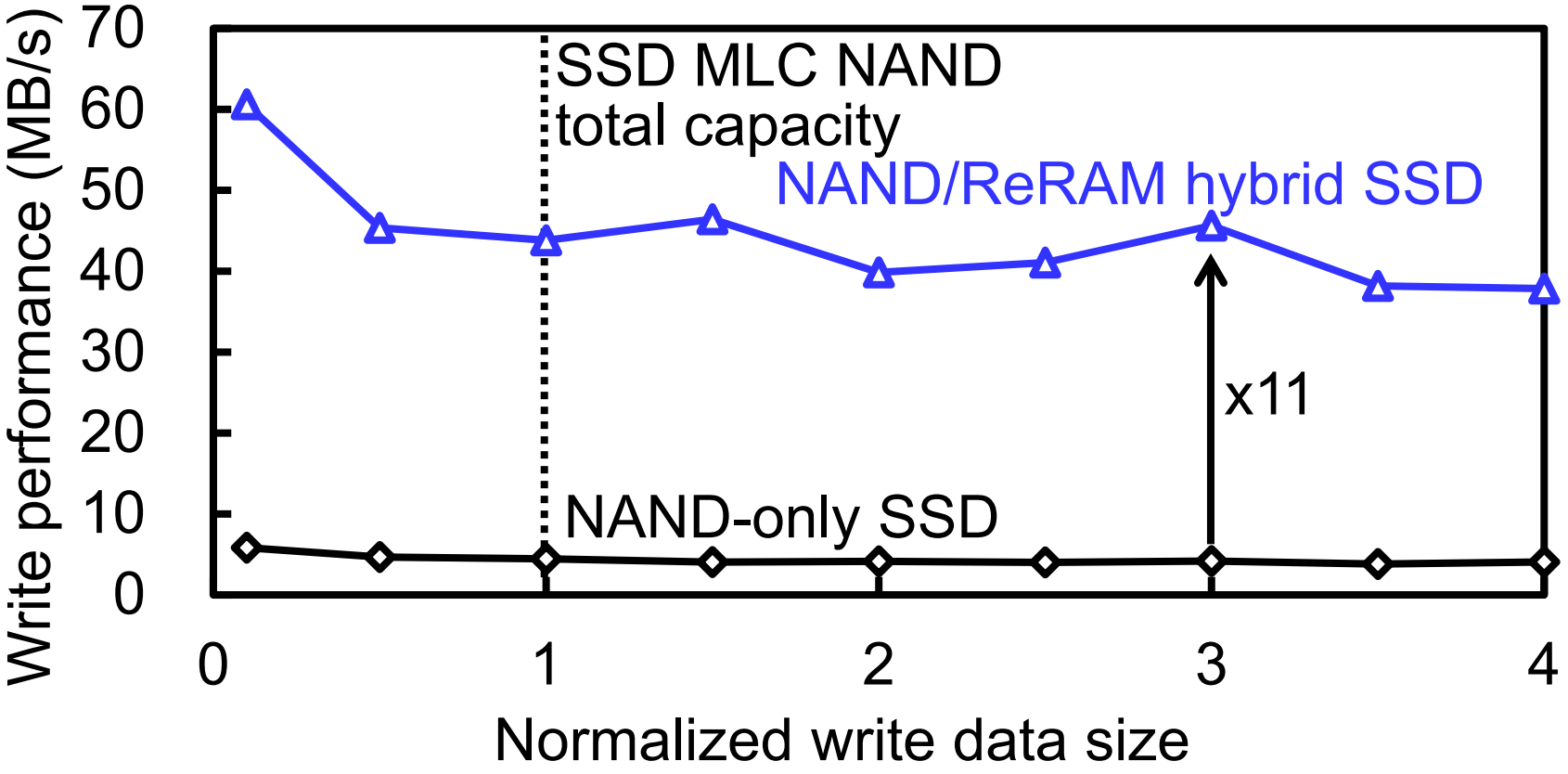
NAND/ReRAM Hybrid Storage (1)



H. Fujii *et al.*, *Symp. VLSI Circ.*, pp. 134-135, 2013.

NAND/ReRAM Hybrid Storage (2)

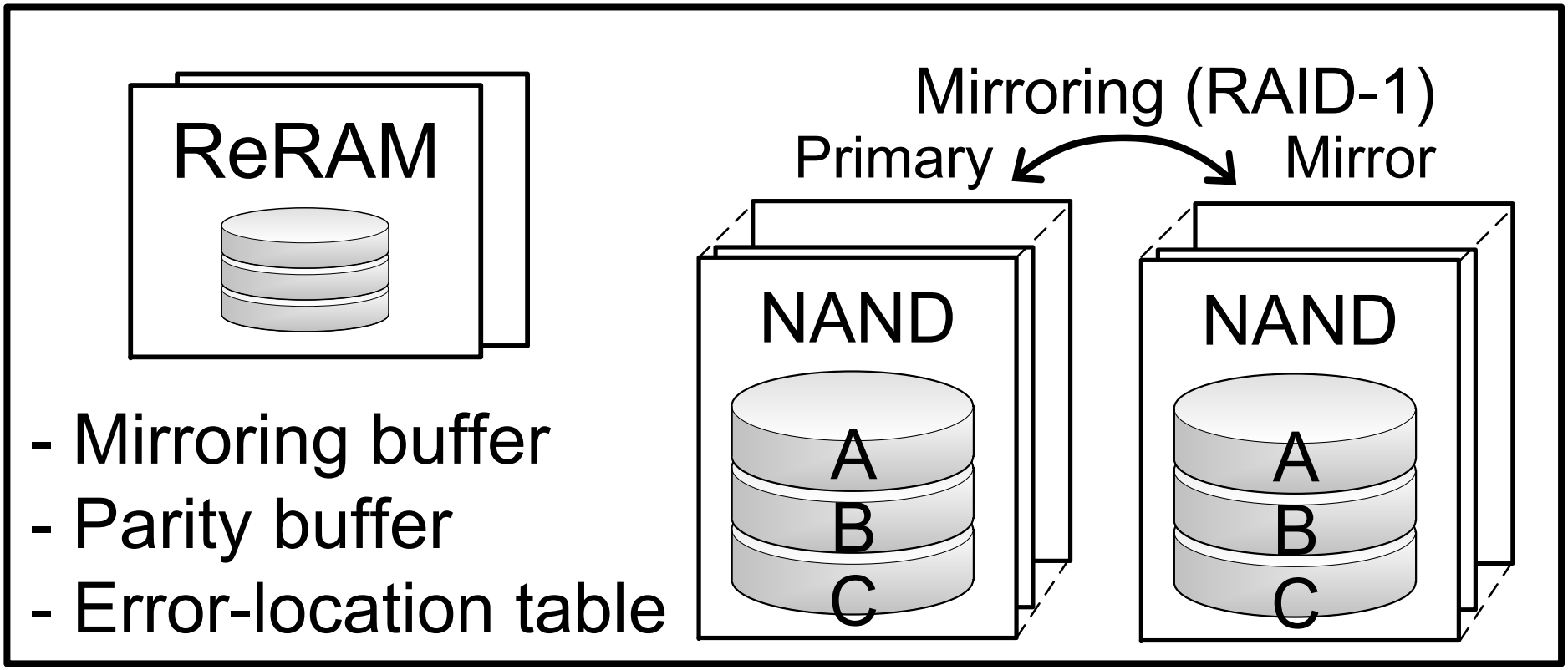
- NAND/ReRAM hybrid storage can boost storage performance by 11x with 6.9x lower write/erase cycle and 93% lower energy compared with NAND only storage.



H. Fujii *et al.*, *Symp. VLSI Circ.*, pp. 134-135, 2013.

Architecture of the Conventional Work

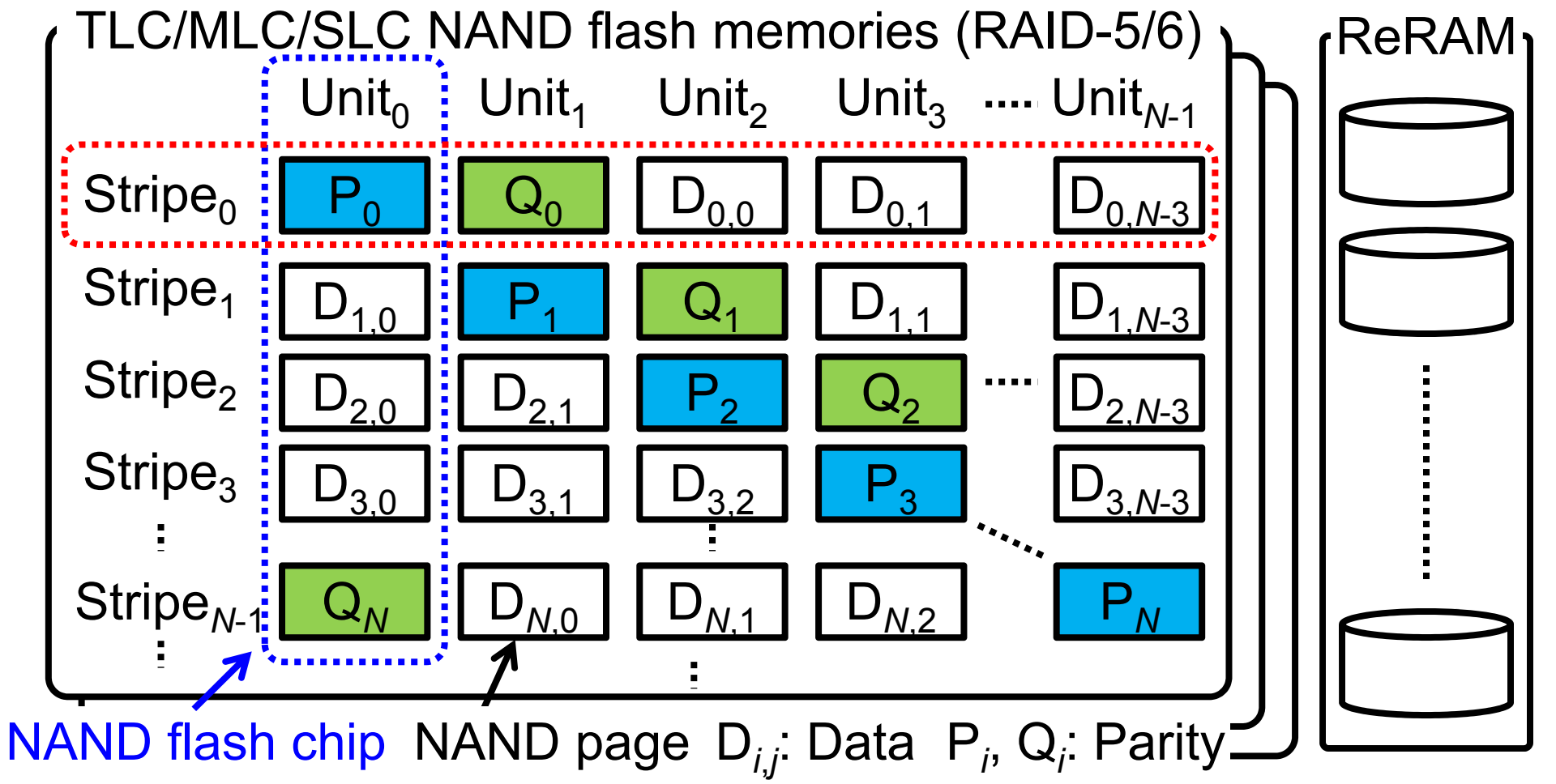
- Data in NAND flash memories are mirrored to secure the data (*).
- The overhead of mirroring is huge (100%).



(*) S. Tanakamaru *et al.*, *ISSCC*, pp. 226-227, 2013.

Architecture of Proposed Storage

- Cost efficient RAID-5/6 is applied to each stripe.
- ReRAM is also used as storage to boost performance.



Comparison of Conv. and Prop. Works

	NAND type	RAID type	Overhead	ReRAM usage
Conv. (*)	MLC (2bit/cell)	RAID-1 (Mirroring)	100%	Parity buffer
Prop.	TLC (3bit/cell)	RAID-5	1/*N* (5% if *N* = 20)	Storage
Prop.	TLC (3bit/cell)	RAID-6	2/*N* (10% if *N* = 20)	Storage

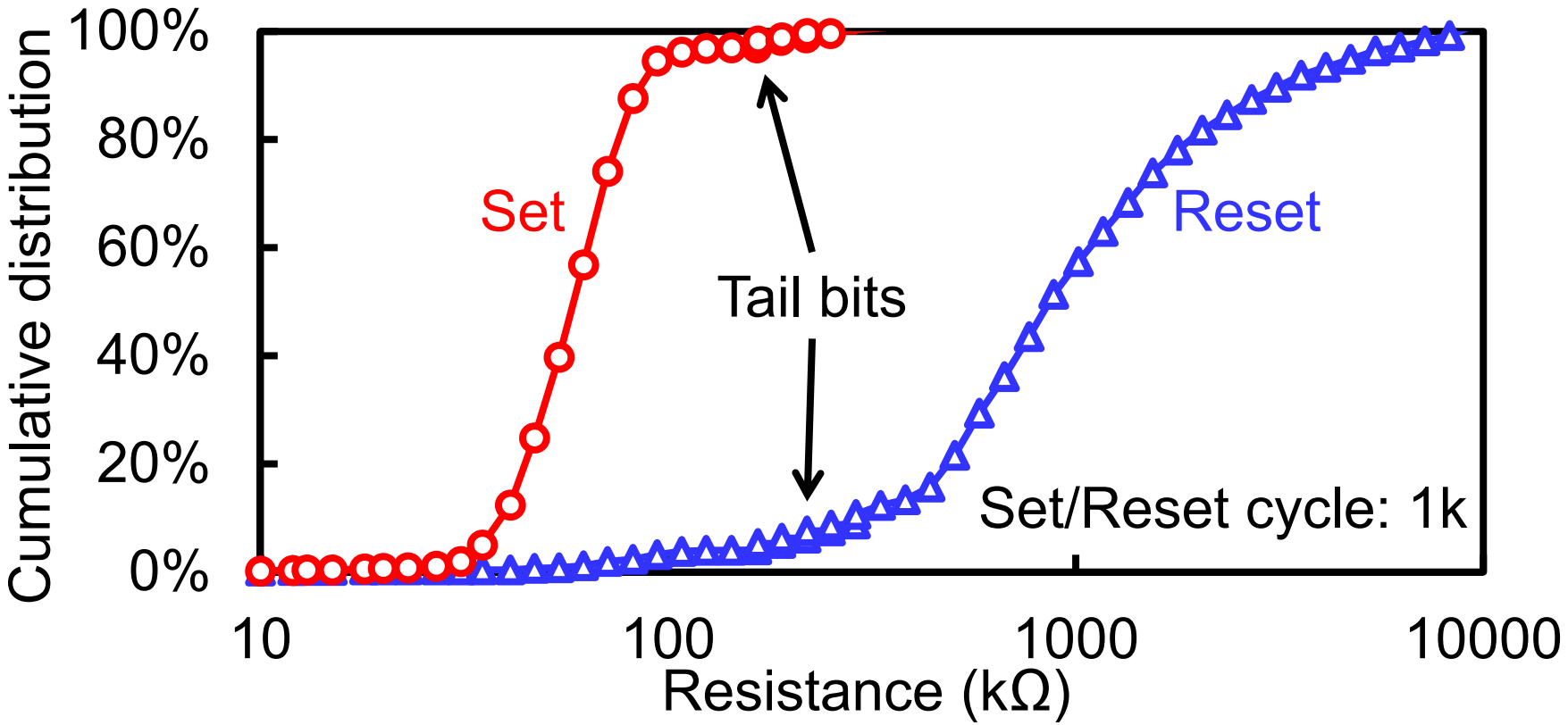
Optimized for NAND flash memory in this work.

(*) S. Tanakamaru *et al.*, *ISSCC*, pp. 226-227, 2013.

Reliability of ReRAM

- Reliability of ReRAM is an issue due to the large resistance variation.

⇨ Needs for highly reliable techniques for ReRAM.

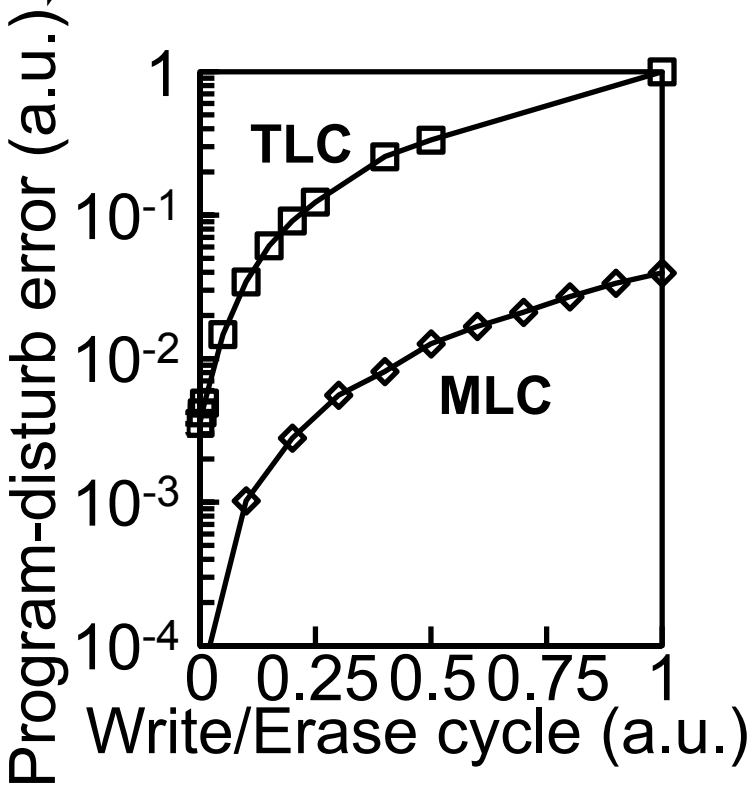


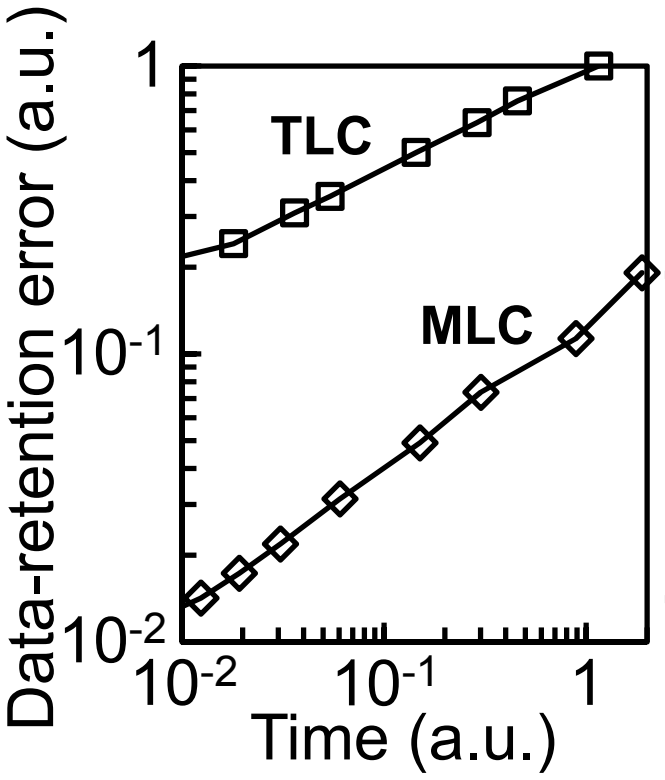
S.Y. Ning *et al.*, *SSDM*, pp. 572-573, 2013.

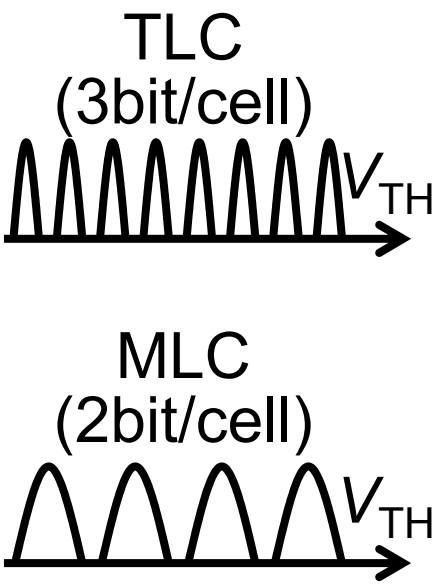
Reliability of TLC NAND Flash Memory

- BER of triple-level cell (TLC) NAND is about 10-times larger than multi-level cell (MLC).

⇨ Highly reliable techniques are also required for TLC.





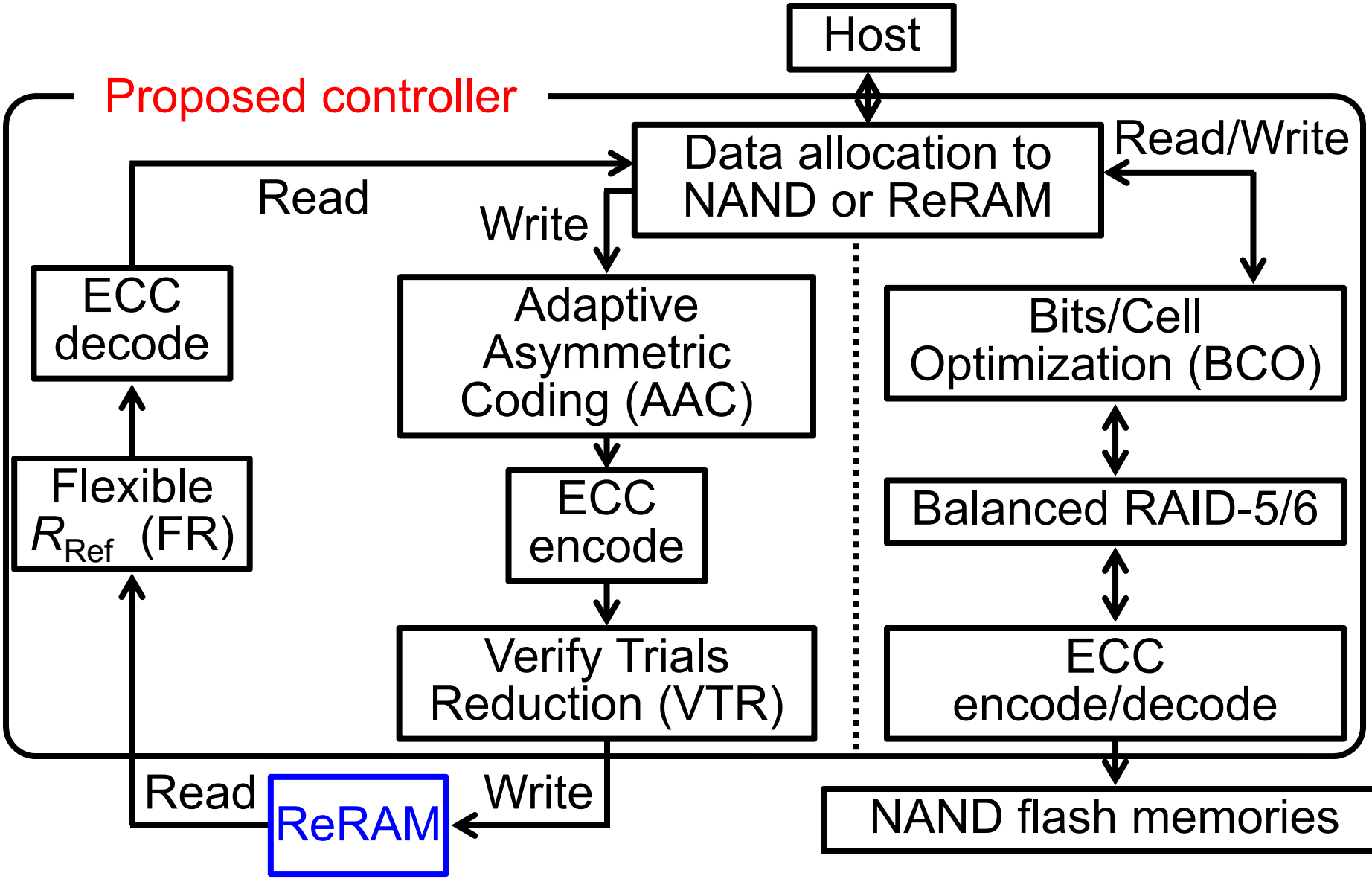


S. Hachiya *et al.*, *SSDM*, pp. 894-895, 2013.

Objectives of This Work

- **ReRAM usage as storage**
  - **Improve the reliability of the ReRAM**
    - **Flexible R_{Ref} (FR)**
    - **Adaptive Asymmetric Coding (AAC)**
  - **Boost the ReRAM performance**
    - **Verify Trials Reduction (VTR)**
- **Improve reliability of TLC NAND based storage**
  - **Optimize lower cost RAID-5/6 for NAND Flash Memories**
    - **Balanced RAID-5/6**
  - **Effective TLC usage**
    - **Bits/Cell Optimization (BCO)**

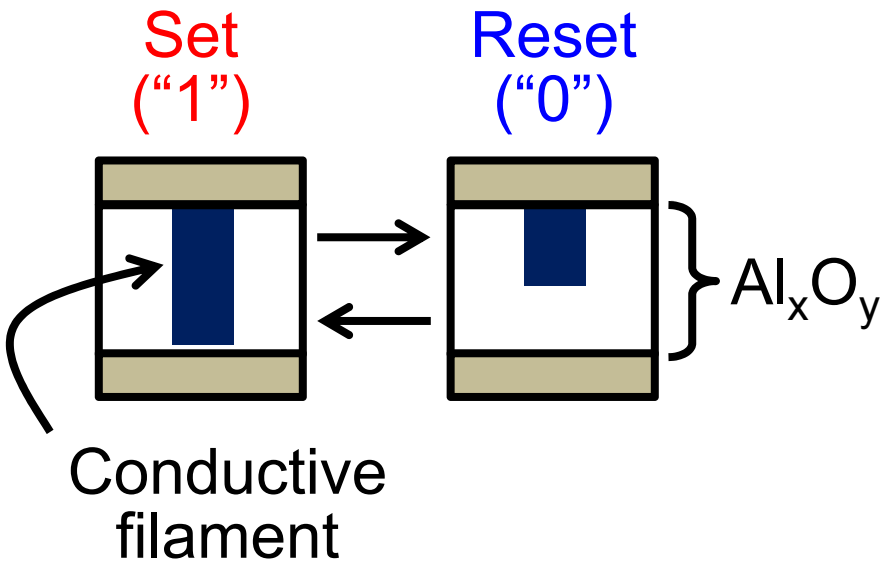
Architecture of the Prop. Controller

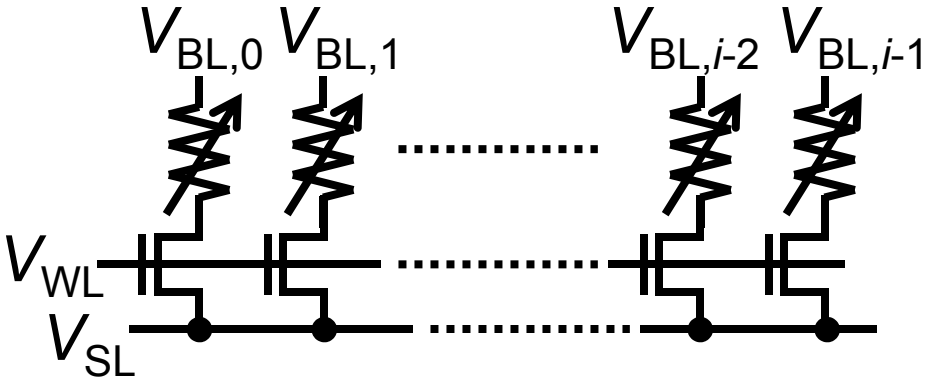


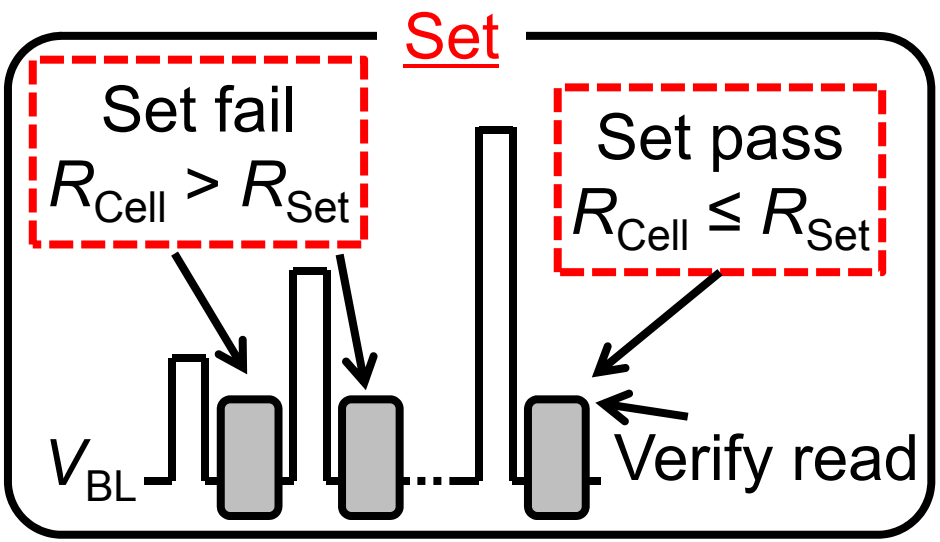
Outline

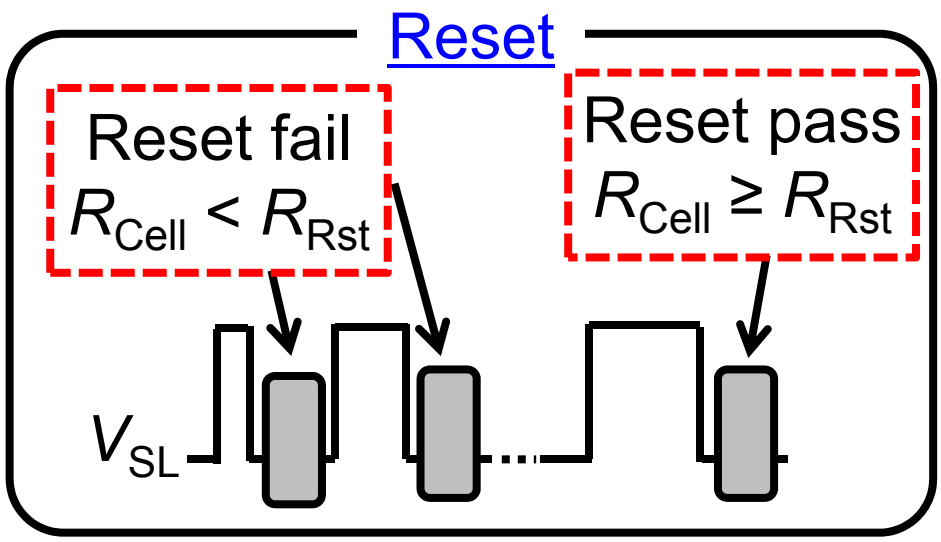
- **Introduction**
- **Reliability Improvement of ReRAM**
  - **Flexible R_{Ref} (FR)**
  - **Adaptive Asymmetric Coding (AAC)**
  - **Verify Trials Reduction (VTR)**
- **Reliability Improvement of NAND Flash Memories**
  - **Balanced RAID-5/6**
  - **Bits/Cell Optimization (BCO)**
- **Summary**

Set/Reset of ReRAM



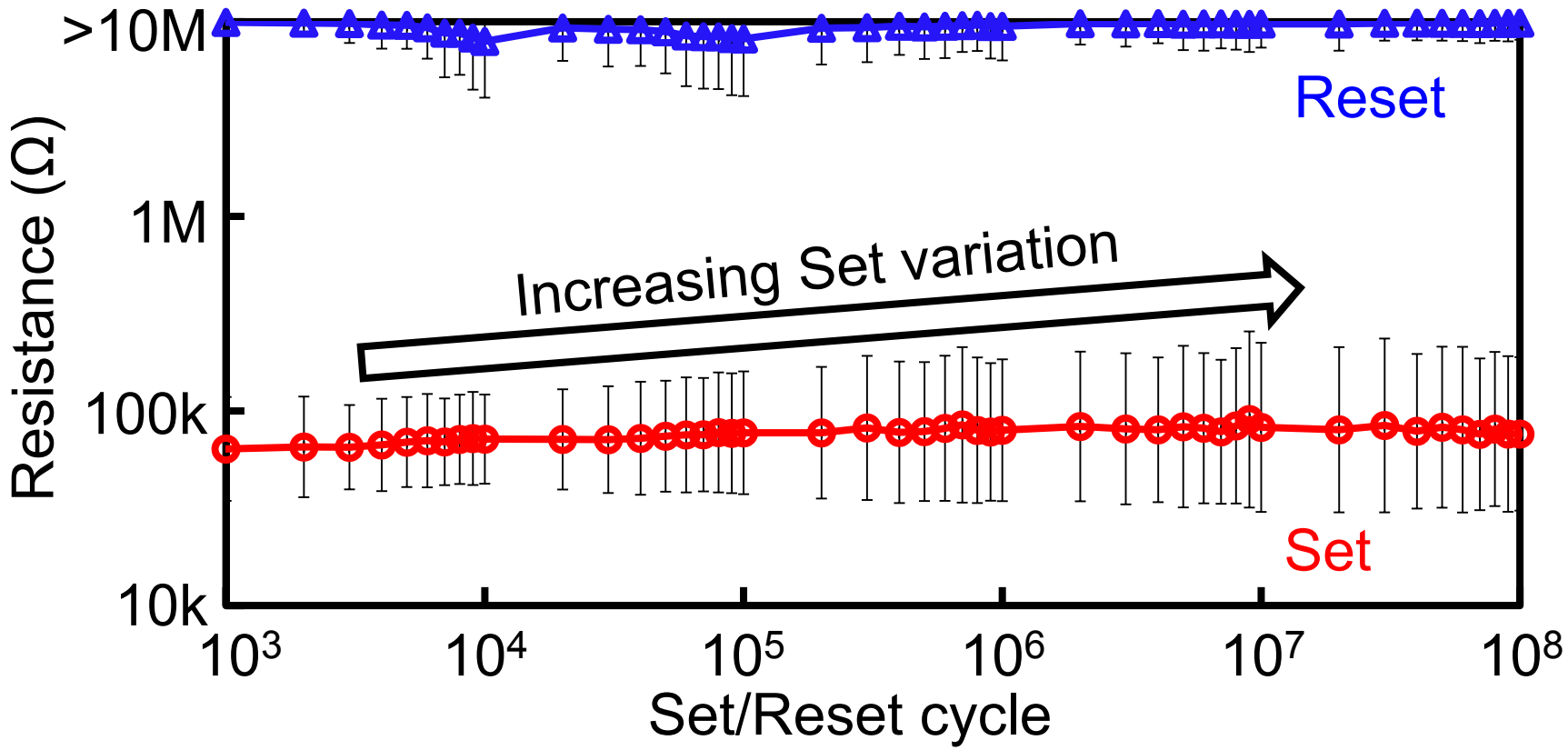






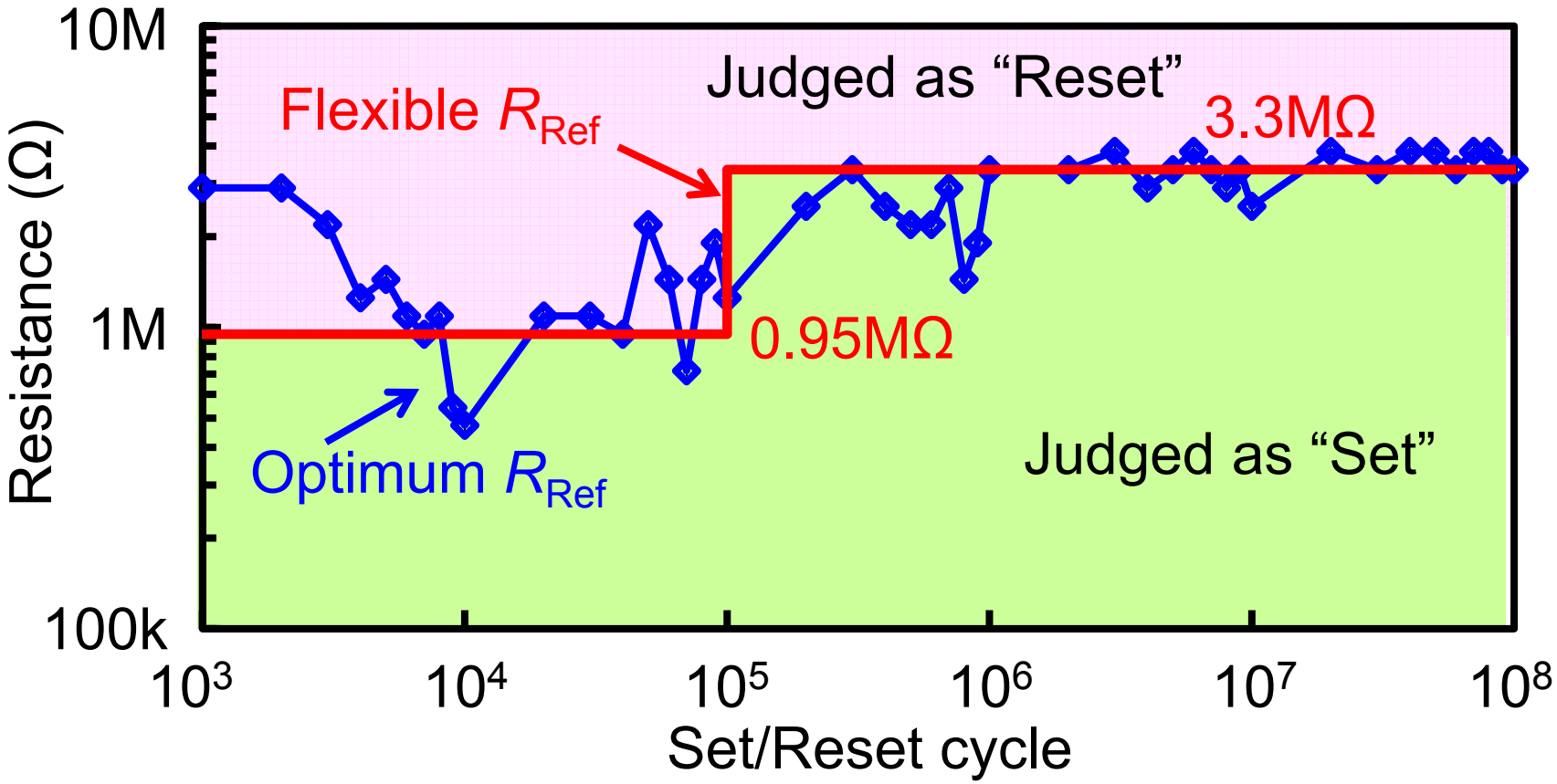
Resistance vs. Set/Reset Cycle

- Set/Reset resistances change during Set/Reset cycle.
- Optimum read reference resistance (R_{Ref}) will also change.



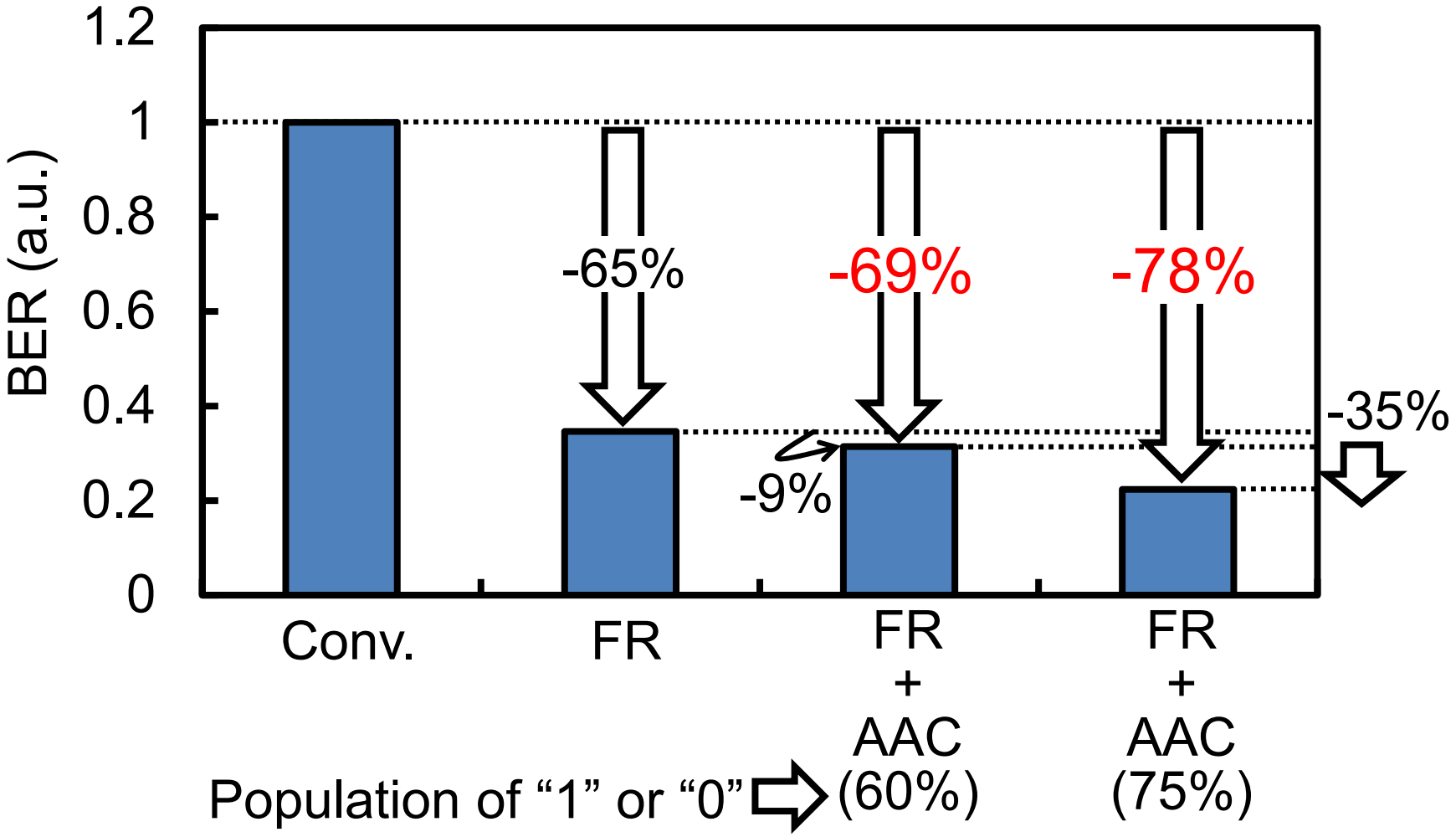
Proposed Flexible R_{Ref} (FR)

- 2-step Flexible R_{Ref} (FR) tracks the optimum R_{Ref}.



BER Improvements

- BER is decreased by 65% by FR and further decreased by Adaptive Asymmetric Coding (AAC).

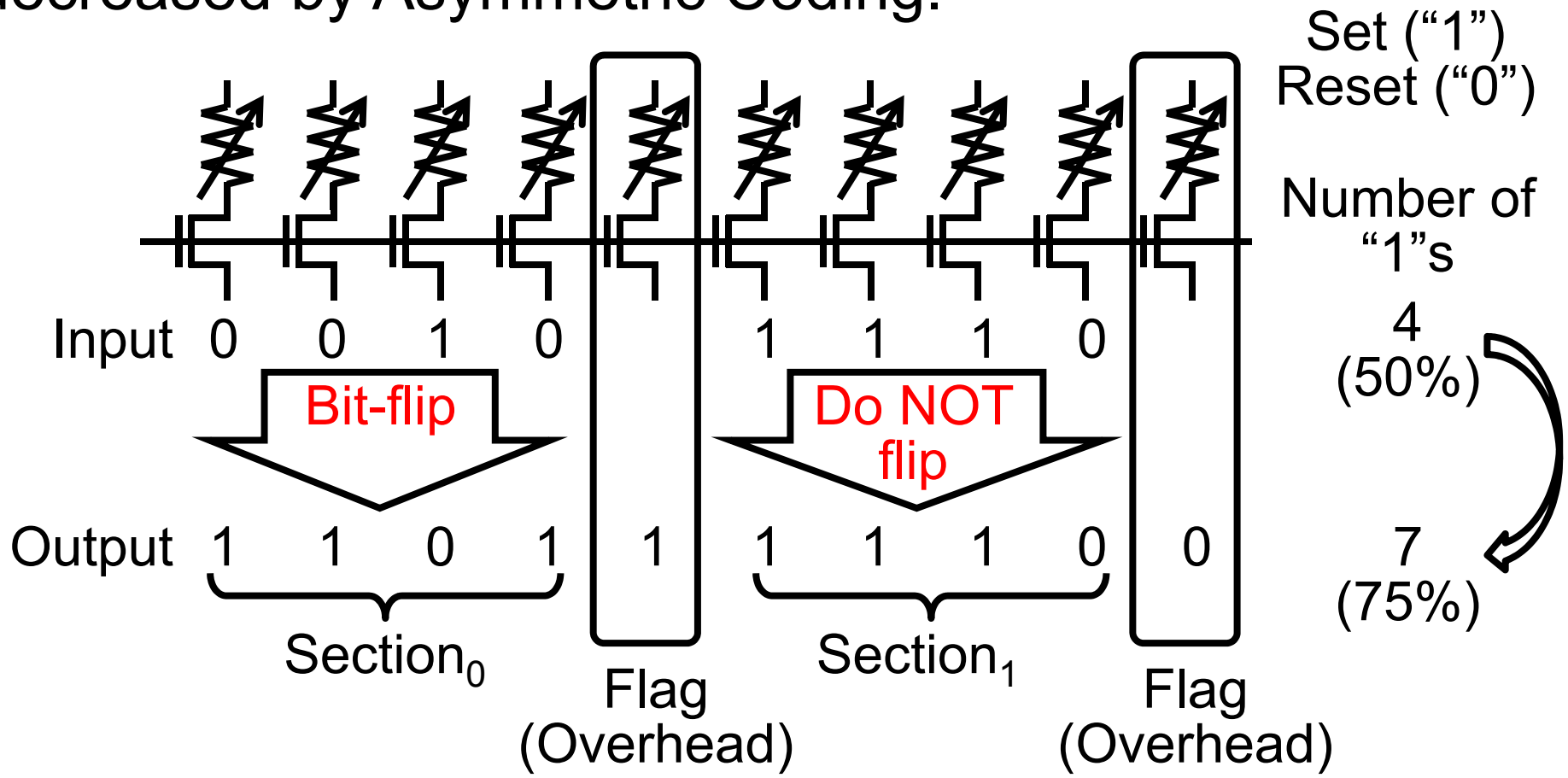


Outline

- **Introduction**
- **Reliability Improvement of ReRAM**
  - **Flexible R_{Ref} (FR)**
  - **Adaptive Asymmetric Coding (AAC)**
  - **Verify Trials Reduction (VTR)**
- **Reliability Improvement of NAND Flash Memories**
  - **Balanced RAID-5/6**
  - **Bits/Cell Optimization (BCO)**
- **Summary**

Asymmetric Coding

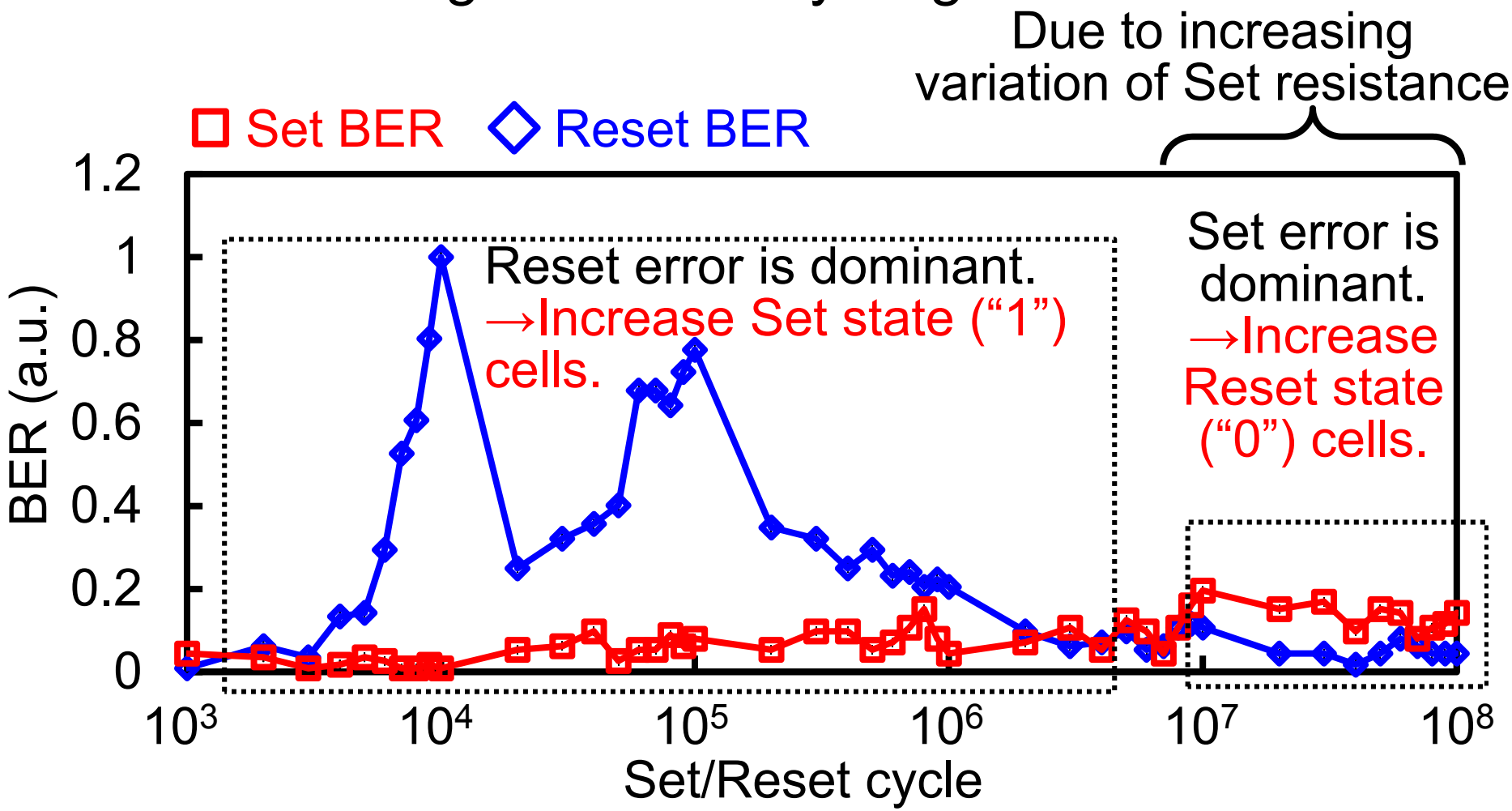
- The population of “1”s (Set) can be increased or decreased by Asymmetric Coding.



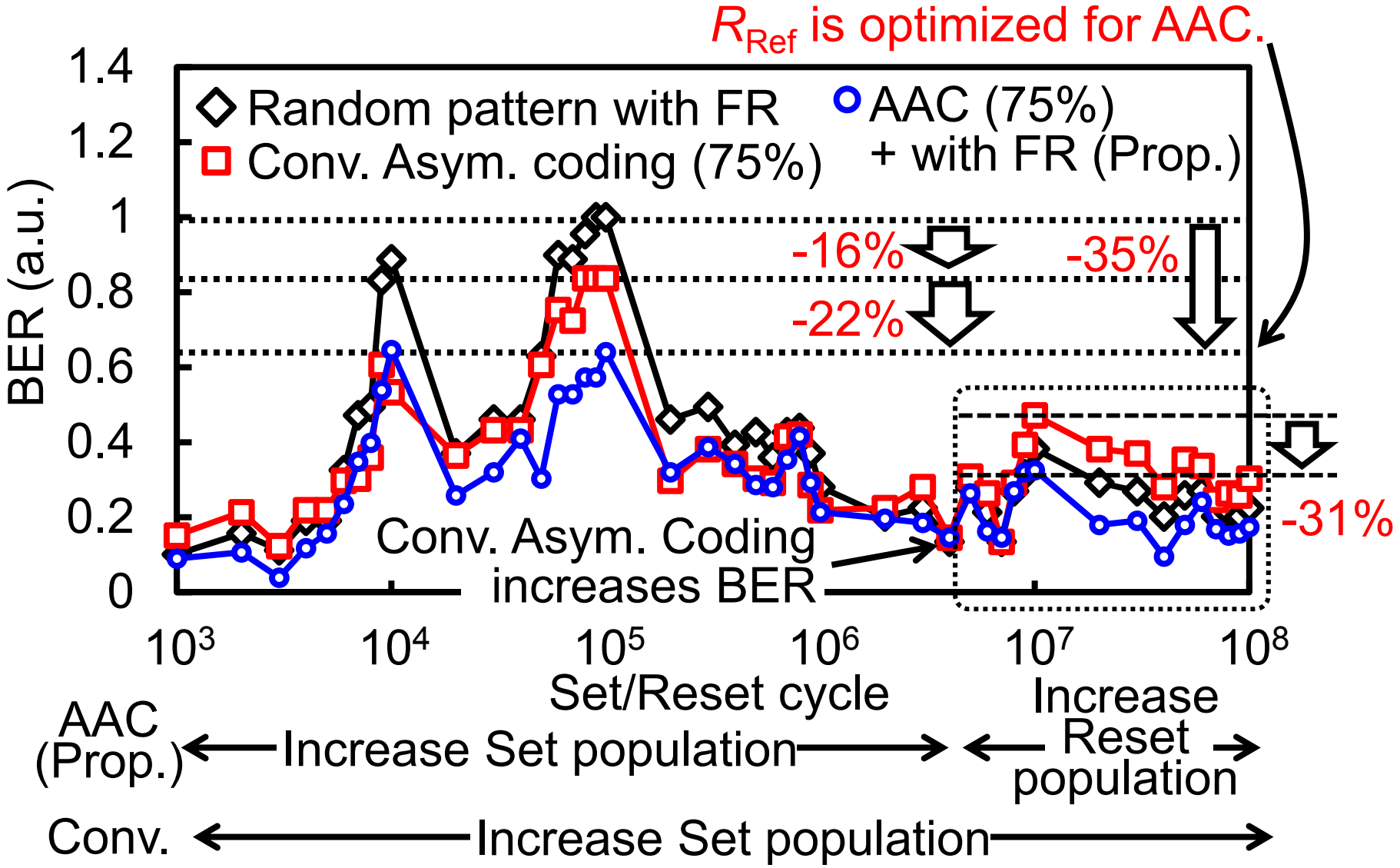
S. Tanakamaru *et al.*, *ISSCC*, pp. 204-205, 2011.

Dominant Error Type

- The dominant error type with FR changes from Reset to Set error during Set/Reset cycling.



Adaptive Asymmetric Coding (AAC)

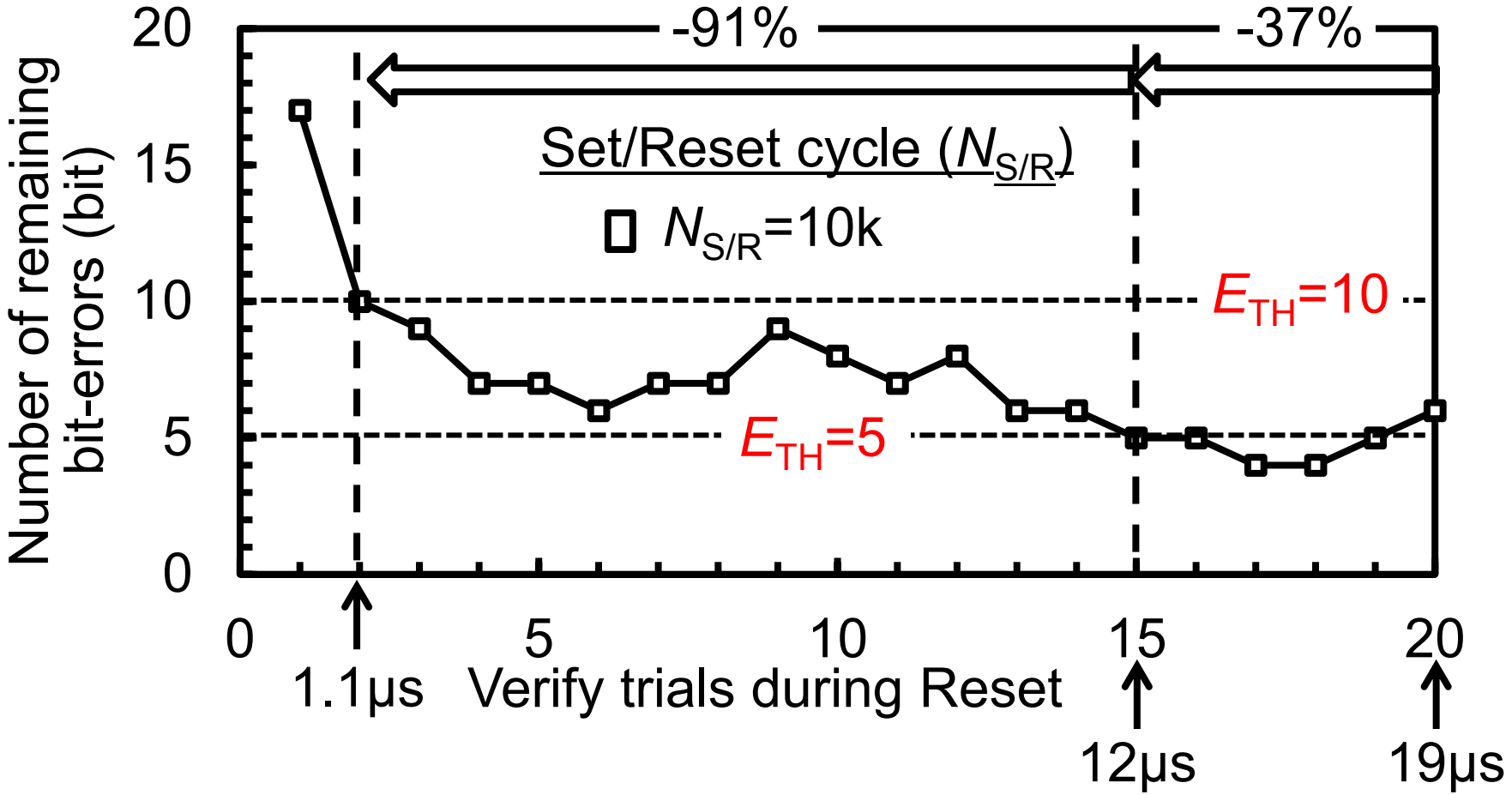


Outline

- **Introduction**
- **Reliability Improvement of ReRAM**
  - **Flexible R_{Ref} (FR)**
  - **Adaptive Asymmetric Coding (AAC)**
  - **Verify Trials Reduction (VTR)**
- **Reliability Improvement of NAND Flash Memories**
  - **Balanced RAID-5/6**
  - **Bits/Cell Optimization (BCO)**
- **Summary**

Measured Remaining Bit-Errors

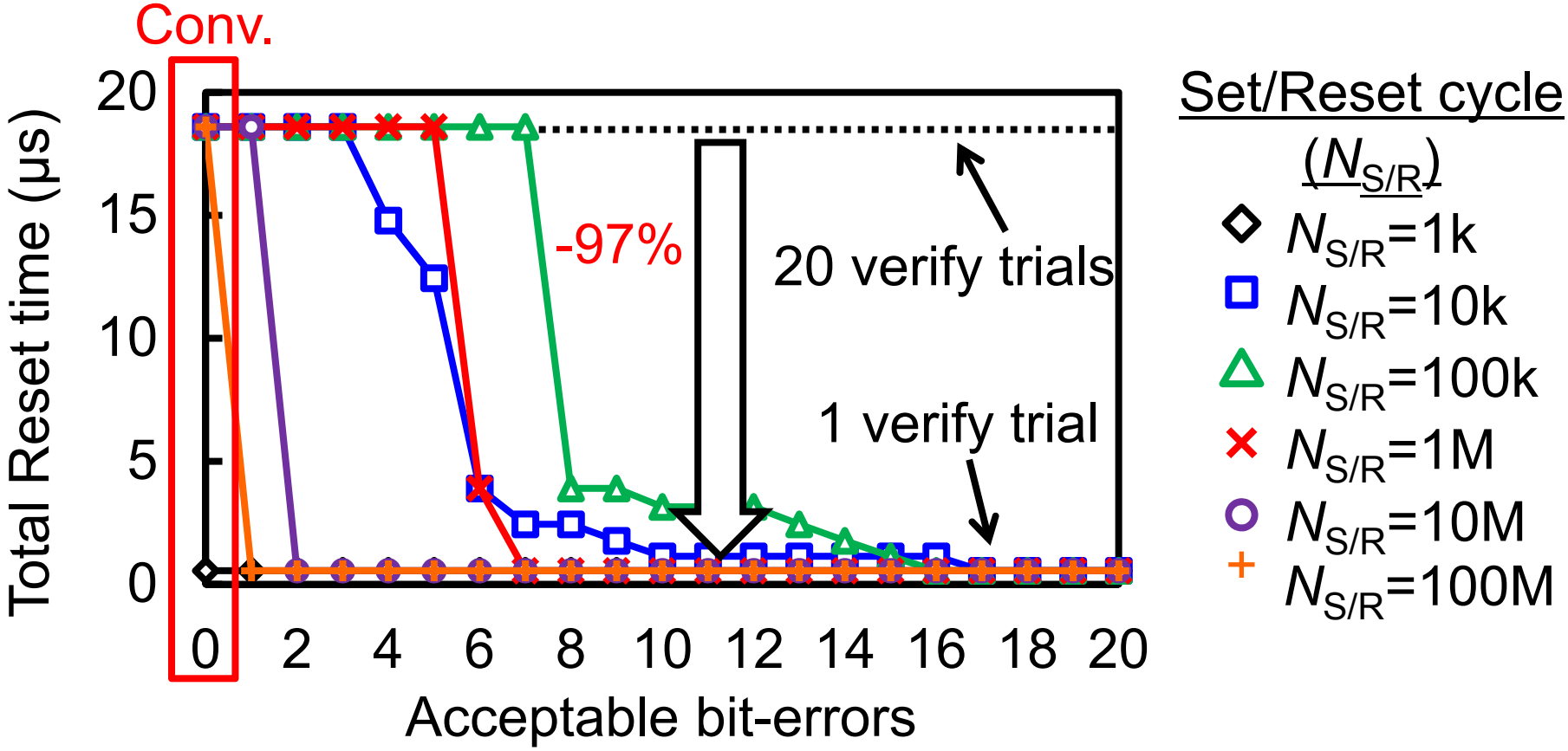
- The number of bit-errors decreases with verify trials.
- ReRAM latency should be below 3μs(*) for high performance.



(*) H. Fujii *et al.*, *Symp. VLSI Circ.*, pp. 134-135, 2013.

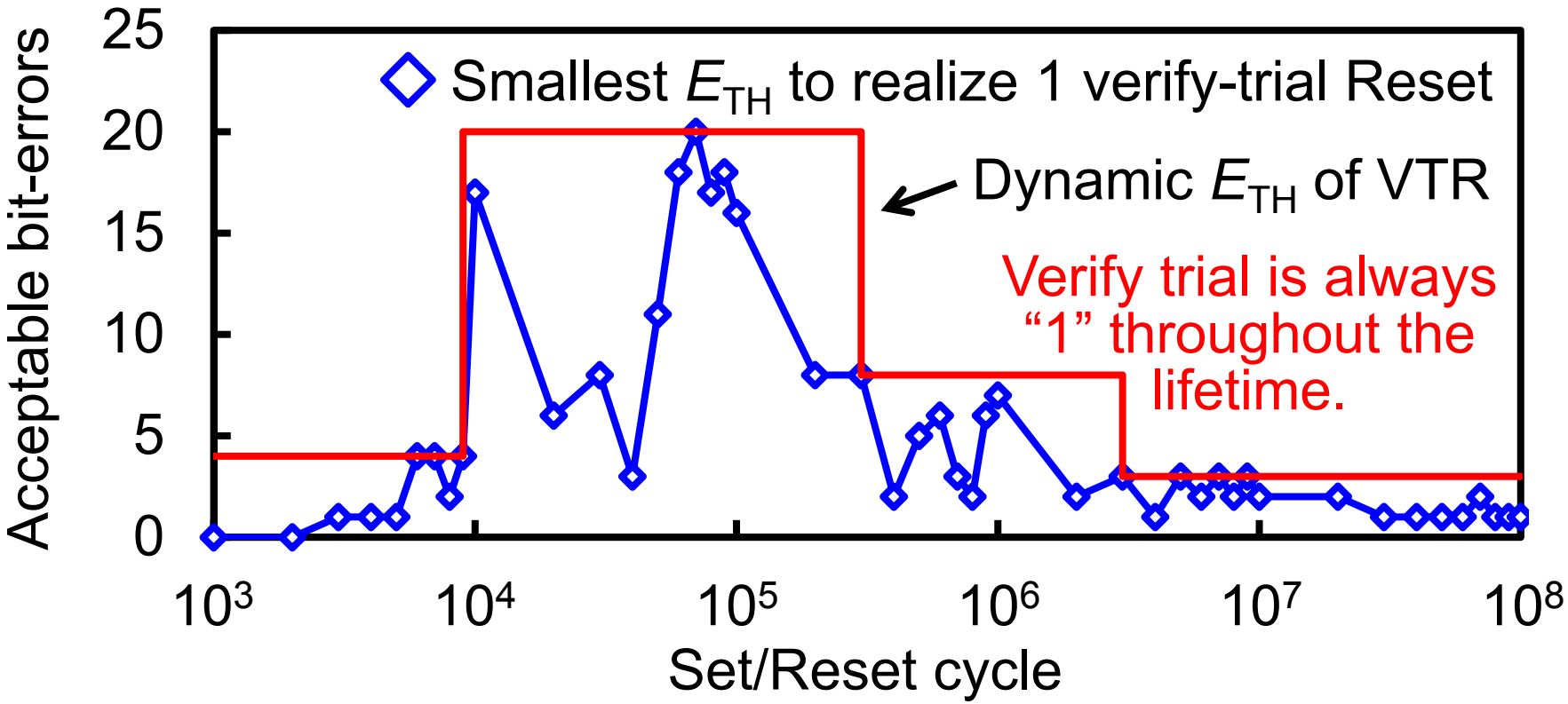
Reset Time Reduction

- Total Reset time can be decreased when bit-errors are accepted during Set/Reset, which are corrected by ECC.



Verify Trials Reduction (VTR)

- By minimizing E_{TH} which can realize 1 verify-trial throughout the lifetime, ECC calculation time (~μs) and power is also minimized.
- E_{TH} (Red line) is determined from the measured minimum E_{TH} (Blue line).



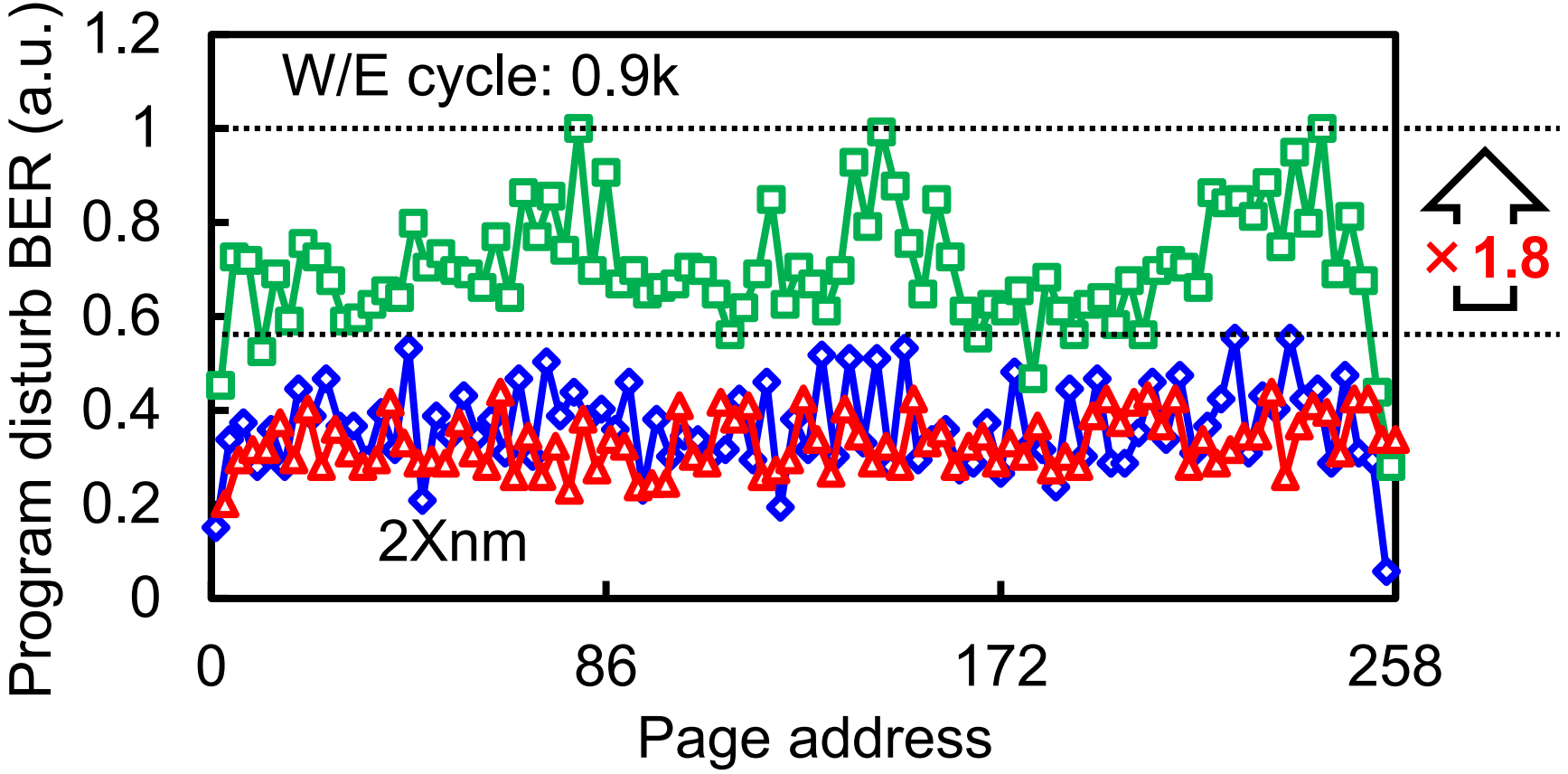
Outline

- **Introduction**
- **Reliability Improvement of ReRAM**
  - **Flexible R_{Ref} (FR)**
  - **Adaptive Asymmetric Coding (AAC)**
  - **Verify Trials Reduction (VTR)**
- **Reliability Improvement of NAND Flash Memories**
  - **Balanced RAID-5/6**
  - **Bits/Cell Optimization (BCO)**
- **Summary**

BER of TLC NAND Flash Memory

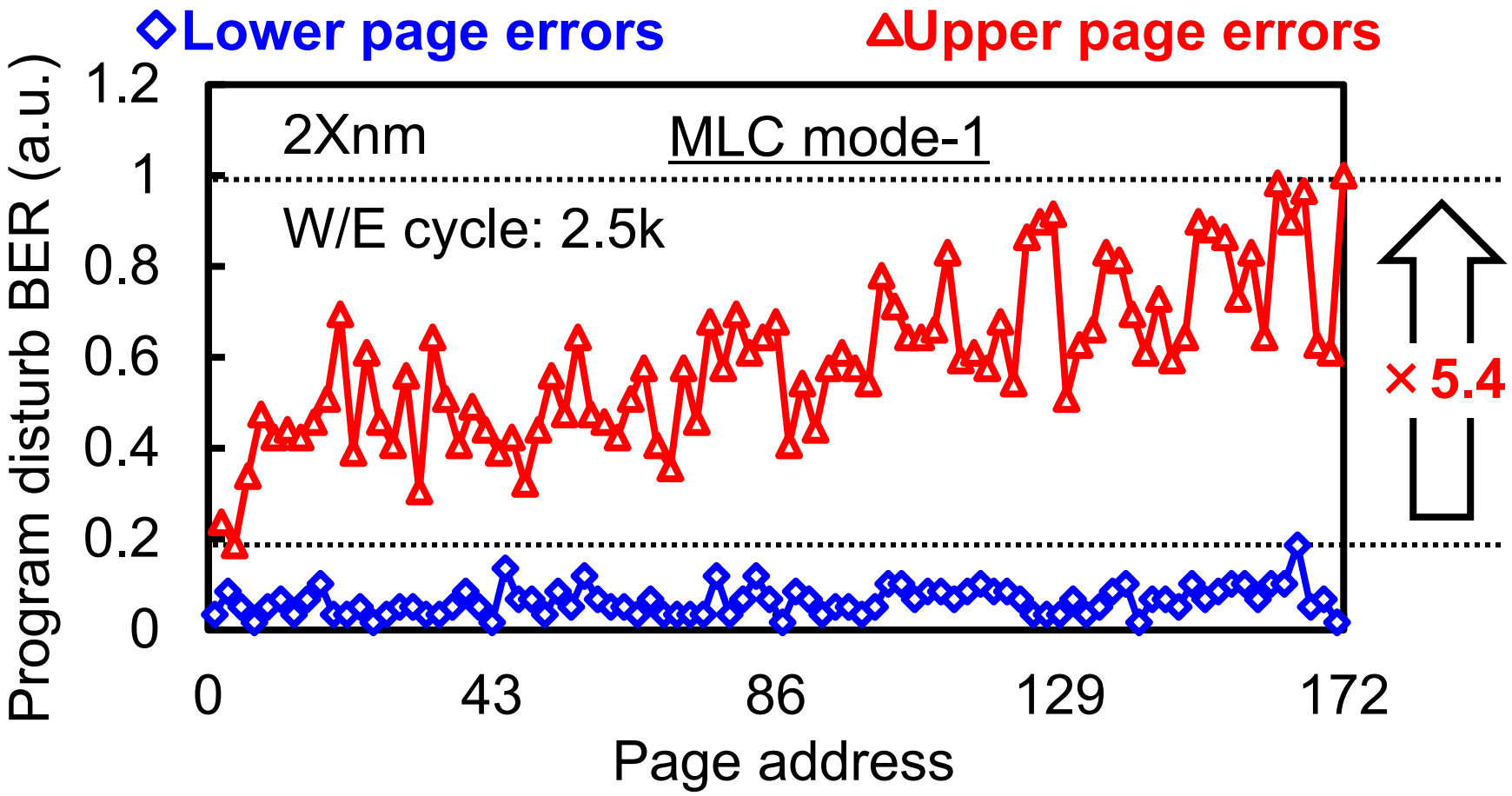
Upper page
Middle page
Lower page

◇Lower page errors □Middle page errors △Upper page errors

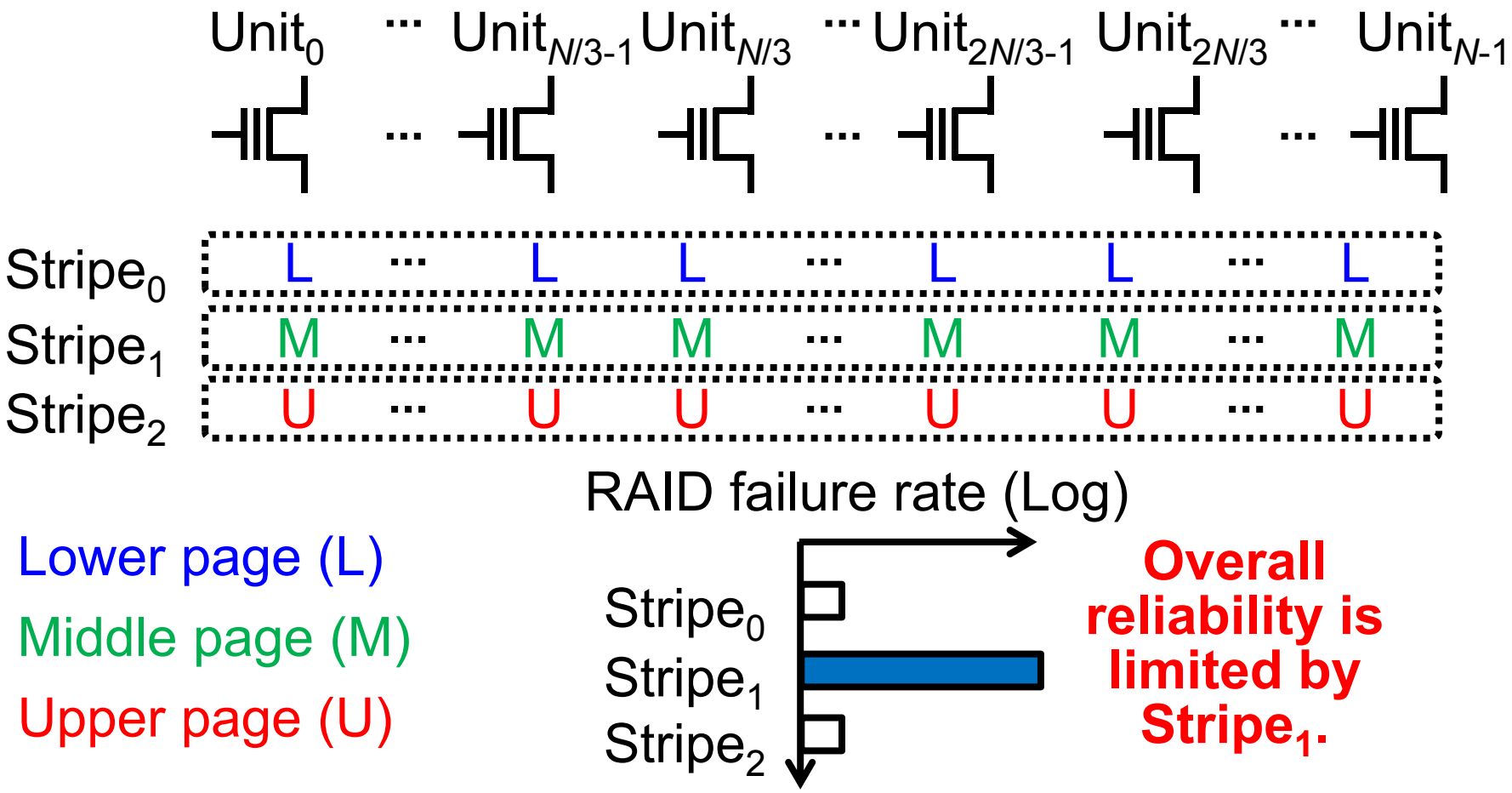


BER of MLC mode-1

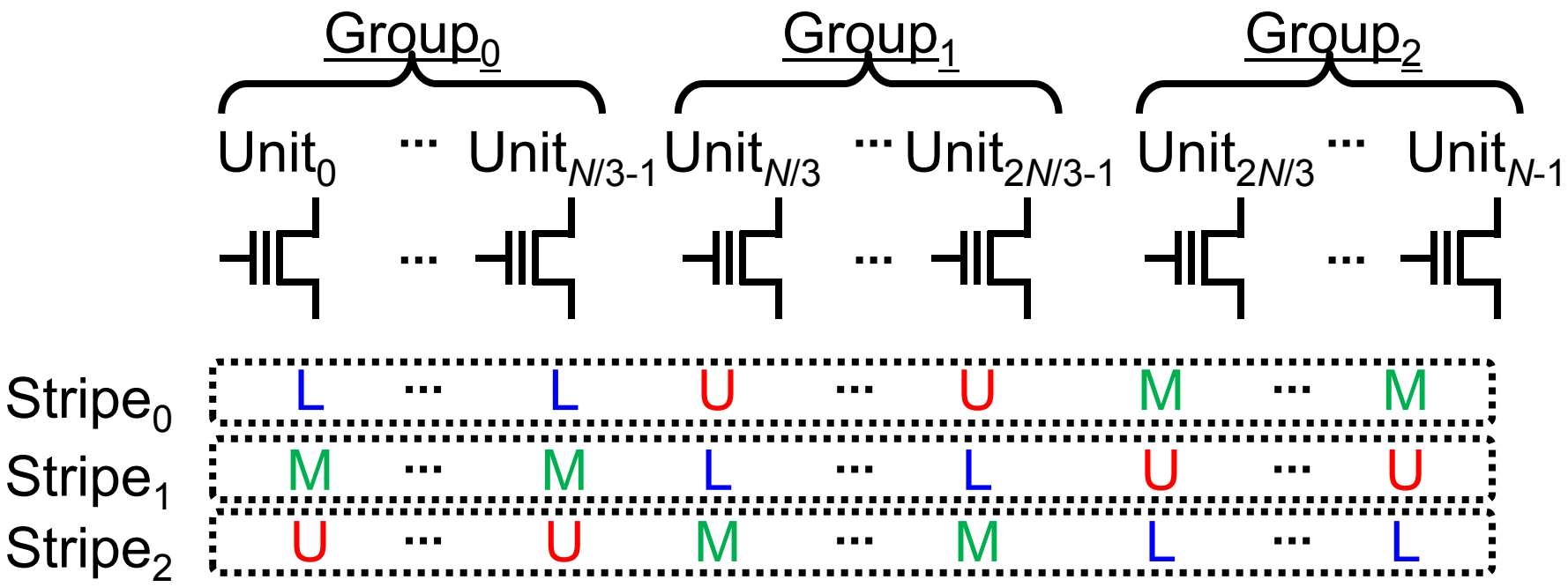
- Upper page BER is higher by 5.4-times.
- BER is strongly dependent on page type.



Conventional RAID-5/6



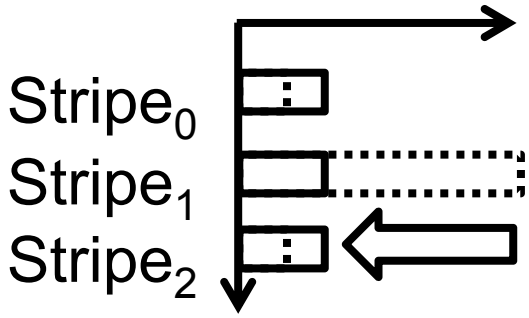
Proposed Balanced RAID-5/6



Lower page (L)

Middle page (M)

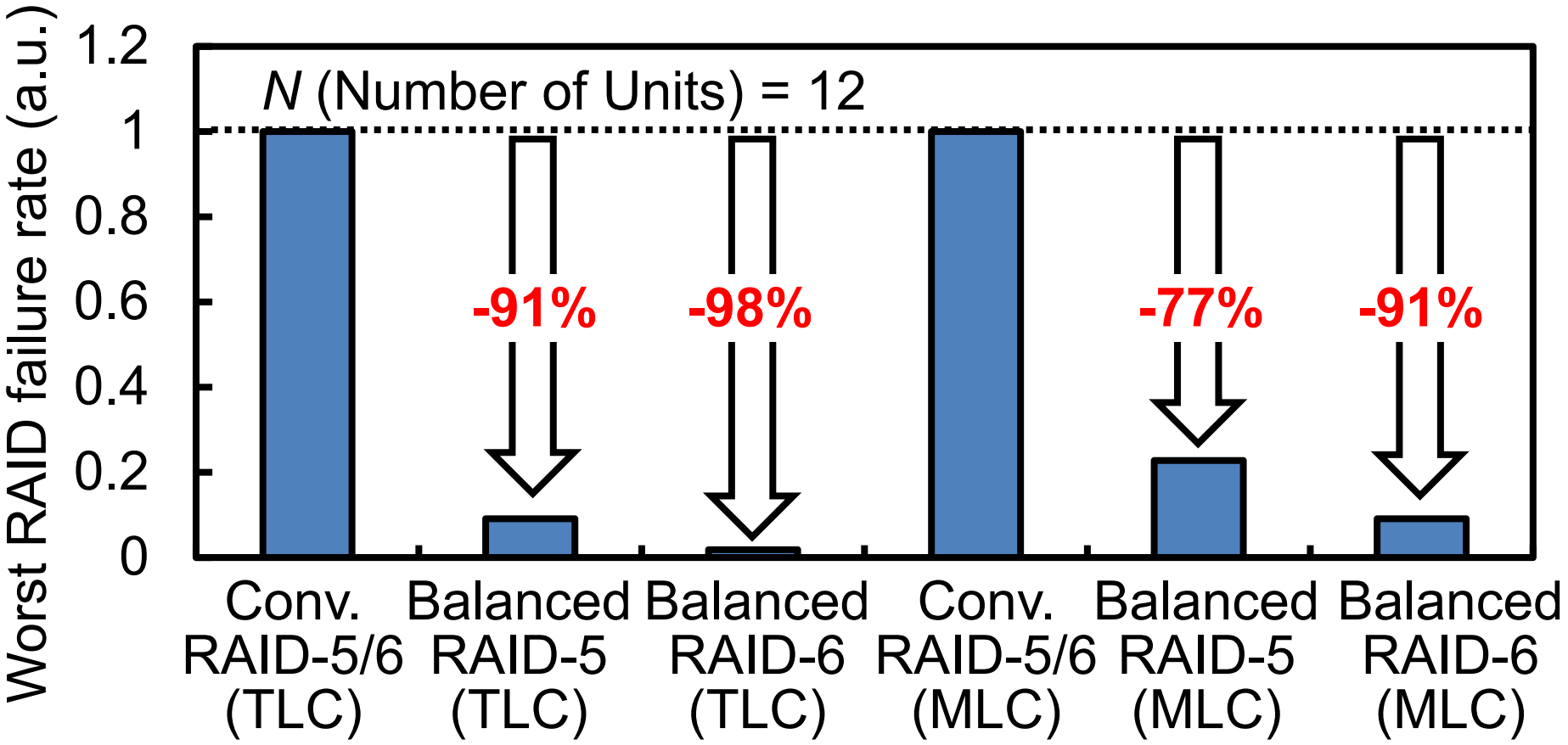
Upper page (U)



RAID failure rate of each stripe is balanced.

Reliability Improvements

- The worst case RAID failure rate improves by 98% with 2Xnm TLC NAND, and in MLC-mode, the improvement is 91%.

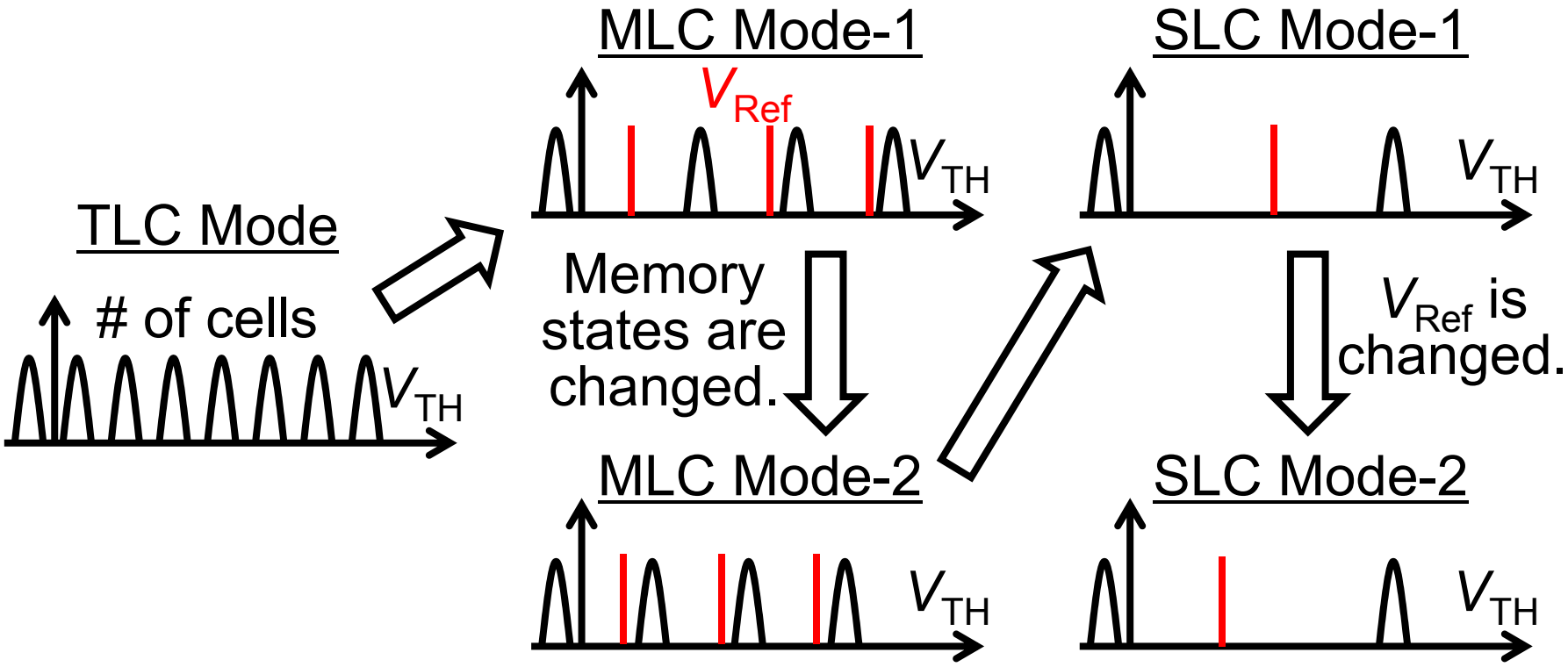


Outline

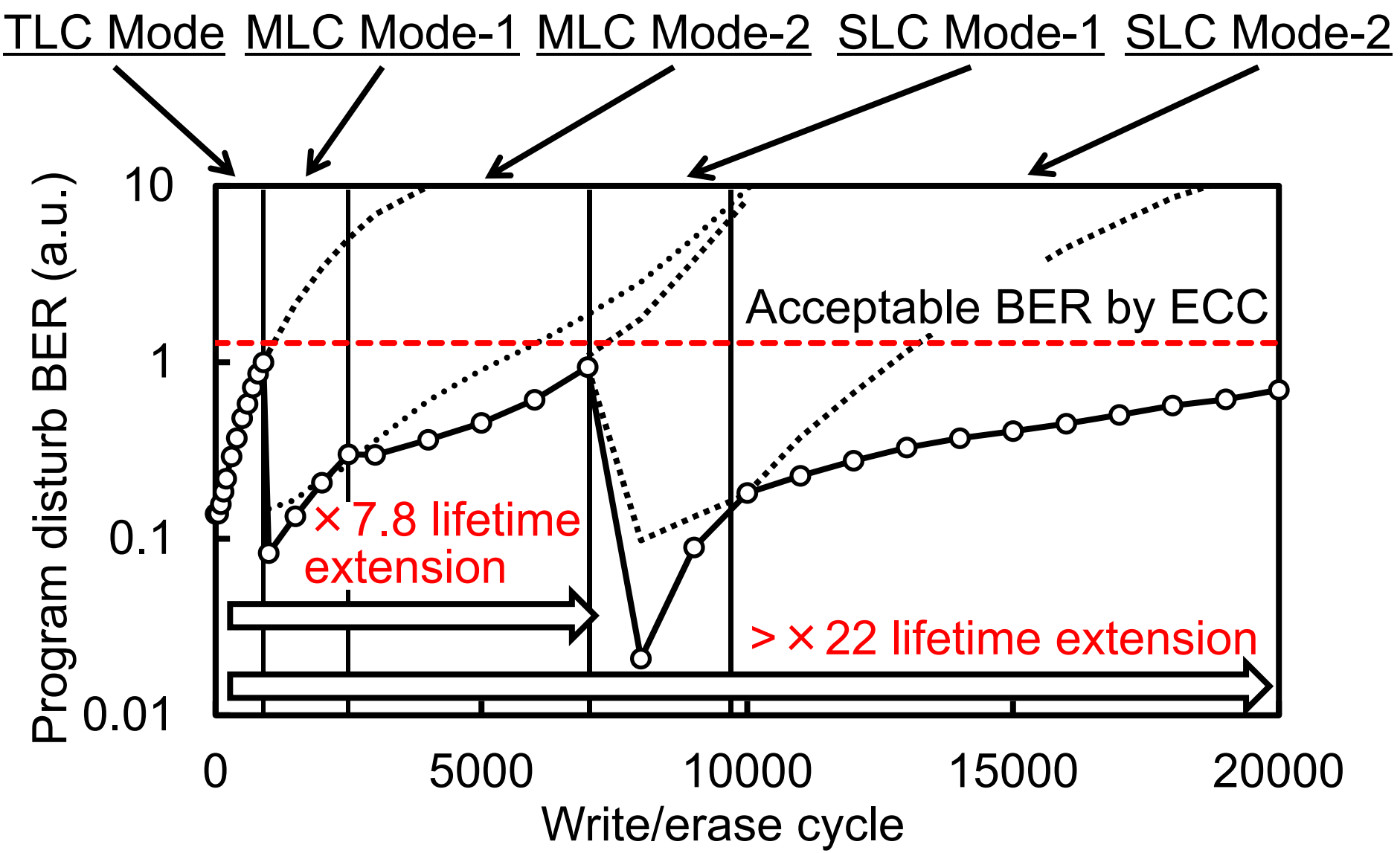
- **Introduction**
- **Reliability Improvement of ReRAM**
  - **Flexible R_{Ref} (FR)**
  - **Adaptive Asymmetric Coding (AAC)**
  - **Verify Trials Reduction (VTR)**
- **Reliability Improvement of NAND Flash Memories**
  - **Balanced RAID-5/6**
  - **Bits/Cell Optimization (BCO)**
- **Summary**

Concept of Bits/Cell Optimization (BCO)

- A worn TLC can be re-allocated as MLC or SLC to extend the chip's useful lifetime.
- Memory states and reference voltage (V_{Ref}) can be optimally selected.



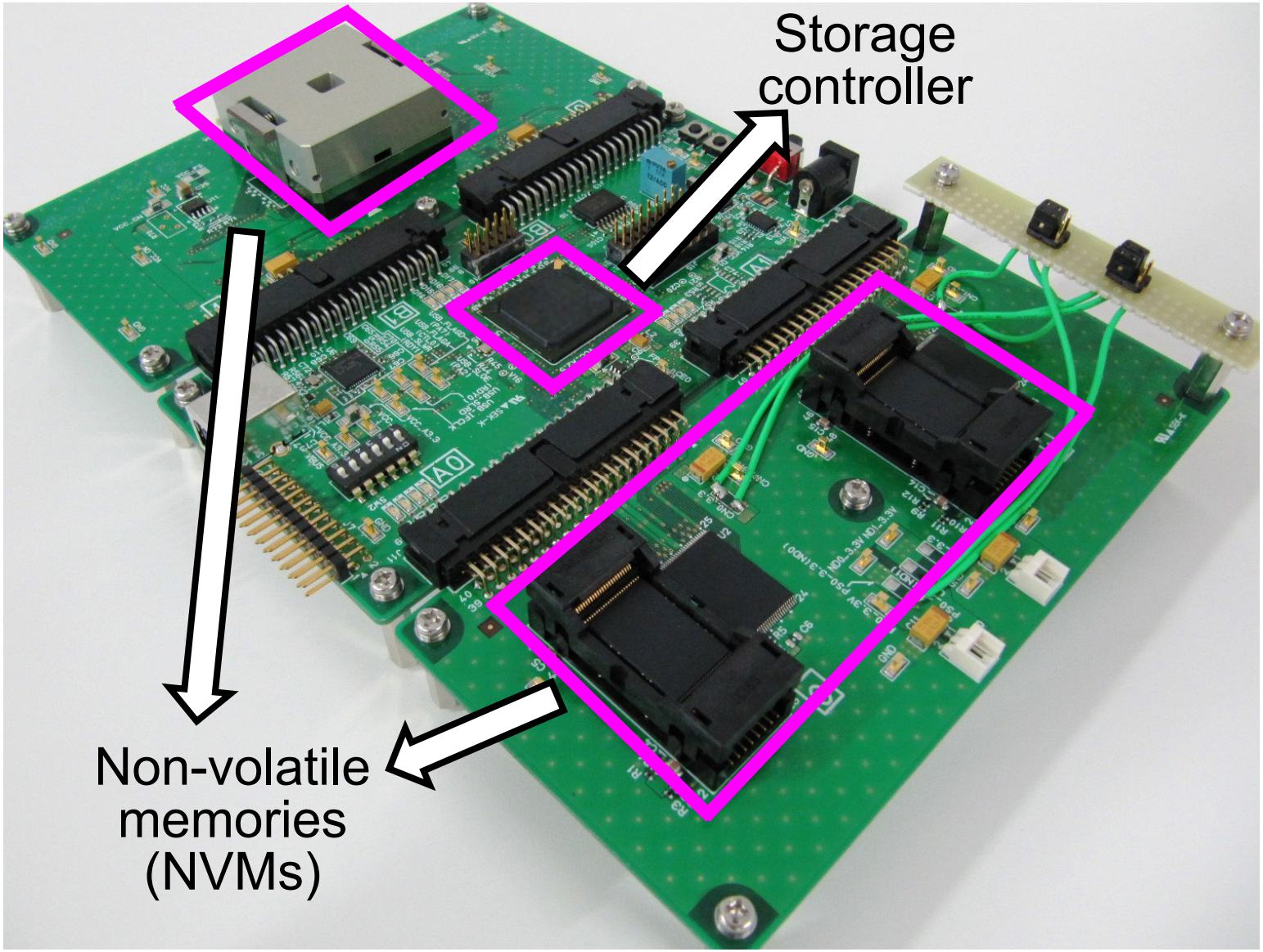
Measured Results of BCO



Outline

- **Introduction**
- **Reliability Improvement of ReRAM**
  - **Flexible R_{Ref} (FR)**
  - **Adaptive Asymmetric Coding (AAC)**
  - **Verify Trials Reduction (VTR)**
- **Reliability Improvement of NAND Flash Memories**
  - **Balanced RAID-5/6**
  - **Bits/Cell Optimization (BCO)**
- **Summary**

Photograph of the Measurement System



Summary of This Work

	Application	Improvement	
Flexible R_{Ref} (FR)	ReRAM	BER -65%	BER -69% (FR & AAC)
Adaptive Asymmetric Coding (AAC)	ReRAM	BER -9%	
Verify Trials Reduction (VTR)	ReRAM	Total Reset time -97%	
Balanced RAID-5/6	TLC NAND flash memory	RAID failure rate -98% (TLC RAID-6)	
Bits/Cell Optimization (BCO)	TLC NAND flash memory	Lifetime (W/E cycle) > × 22	

Conclusion

- **A hybrid storage architecture of ReRAM and TLC NAND flash with RAID-5/6 is developed.**
- **Flexible R_{Ref} (FR) and Adaptive Asymmetric Coding (AAC) reduces BER of ReRAM by 69%.**
- **Verify Trials Reduction (VTR) decreases the program time of ReRAM by 97%.**
- **Balanced RAID-5/6 improves the RAID failure rate by 98%.**
- **Bits/Cell Optimization (BCO) extends the lifetime of NAND by >22x.**

Acknowledgement

This work is partially supported by CREST/JST

Thank you for your attention

A 16Gb ReRAM with 200MB/s Write and 1GB/s Read in 27nm Technology

ISSCC '14

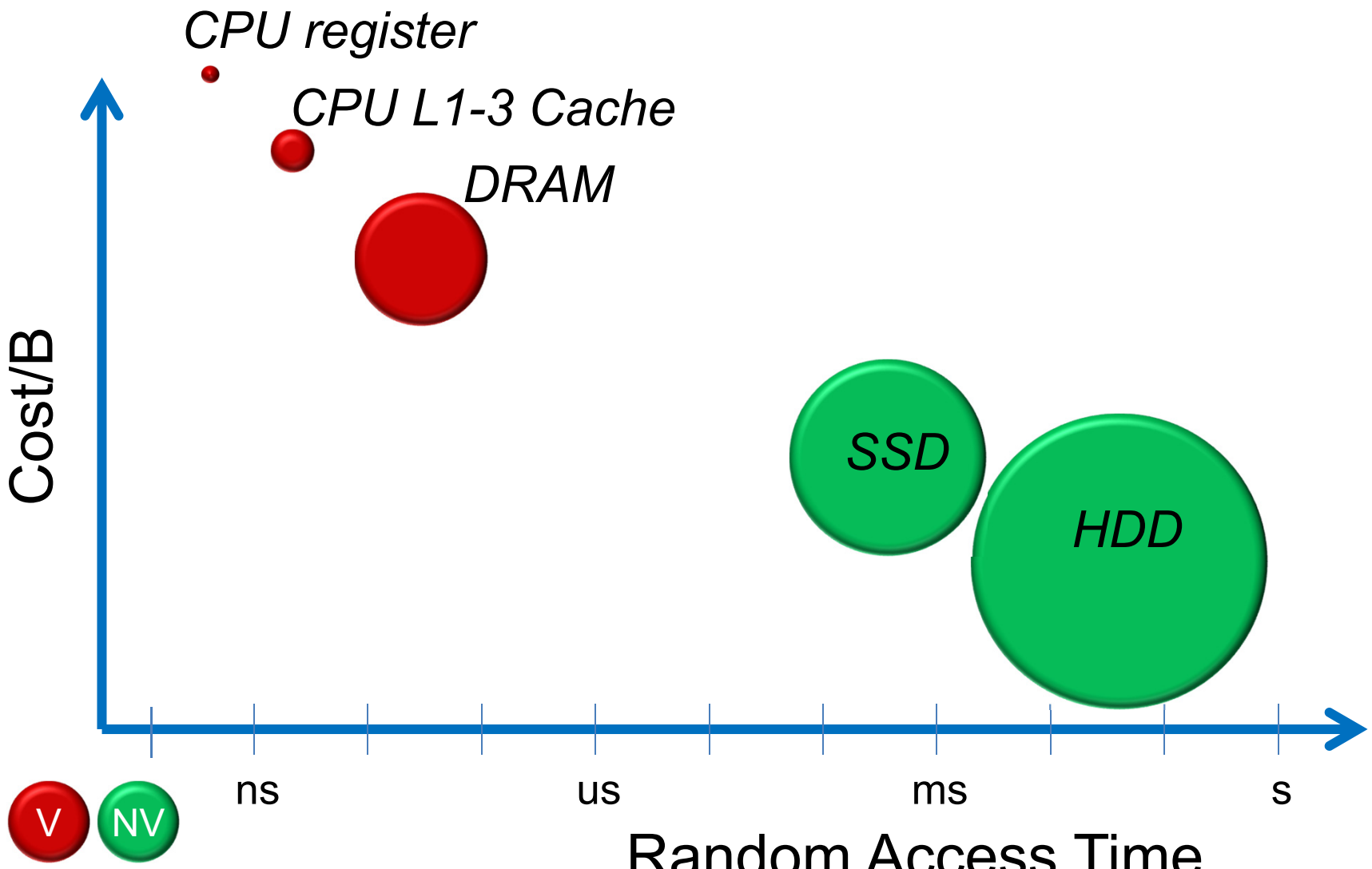
Greg Atwood[1], Rich Fackenthal[1], Mark Fischer[1], Glen Hush[1], Johnny Javanifard[1], Adam Johnson[1], Makoto Kitagawa[2], Yogesh Luthra[1], Chris Martin[1], Duane Mills[1], Yotaro Mori[2], Adachi Naohiro[2], Farid Nemati[1], Wataru Otsuka[2], Mike Pearson[1], Michele Piccardi[1], Kirk Prall[1], Raed Sabbah[1], Lui Sakai[2], Alessandro Sansai[1], Yoshiyuki Shibahara[2], Kerry Tedrow[1], Haruhiko Terada[2], Tomohito Tsushima[2], Keiichi Tsutsui[2], Jack Wu[1]

1. Micron
2. Sony

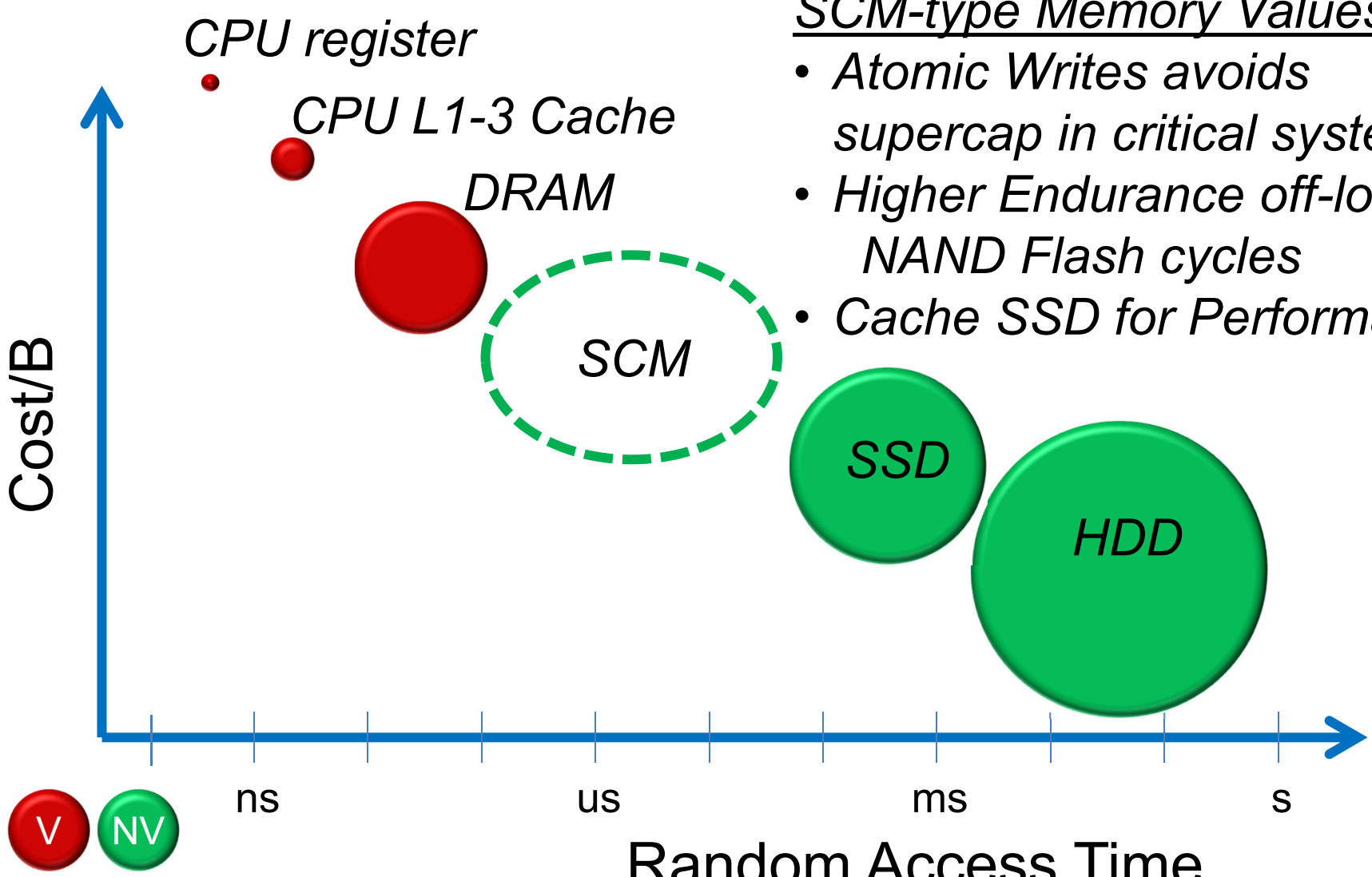
Outline

- Introduction
- Cell
- Array Architecture
- Interface and Data Path
- Sense Amplifier
- Smart Charging
- Column Redundancy
- Silicon Measurements
- Summary Table
- Conclusion

Introduction: The Memory Hierarchy



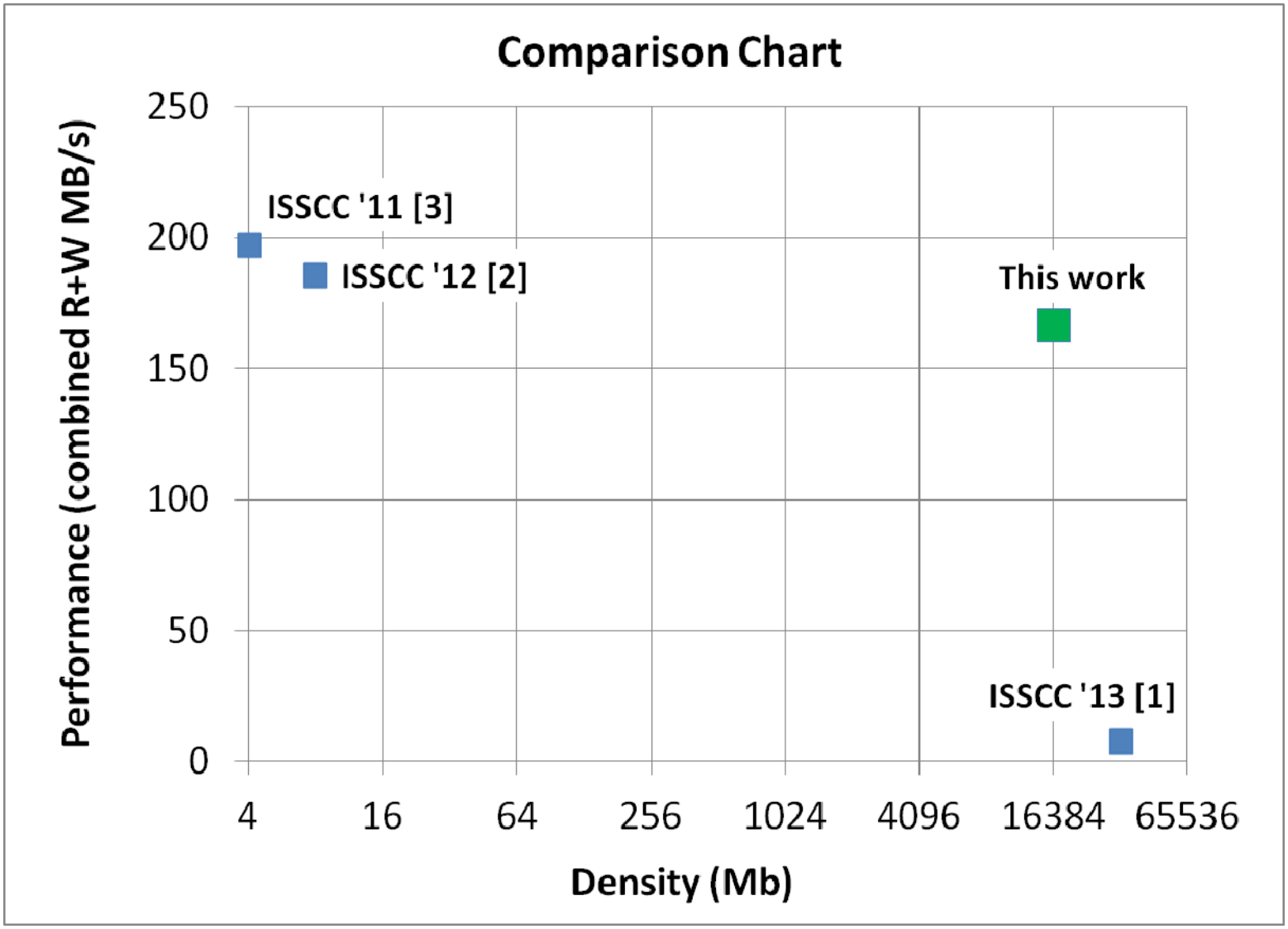
Introduction: The Memory Hierarchy



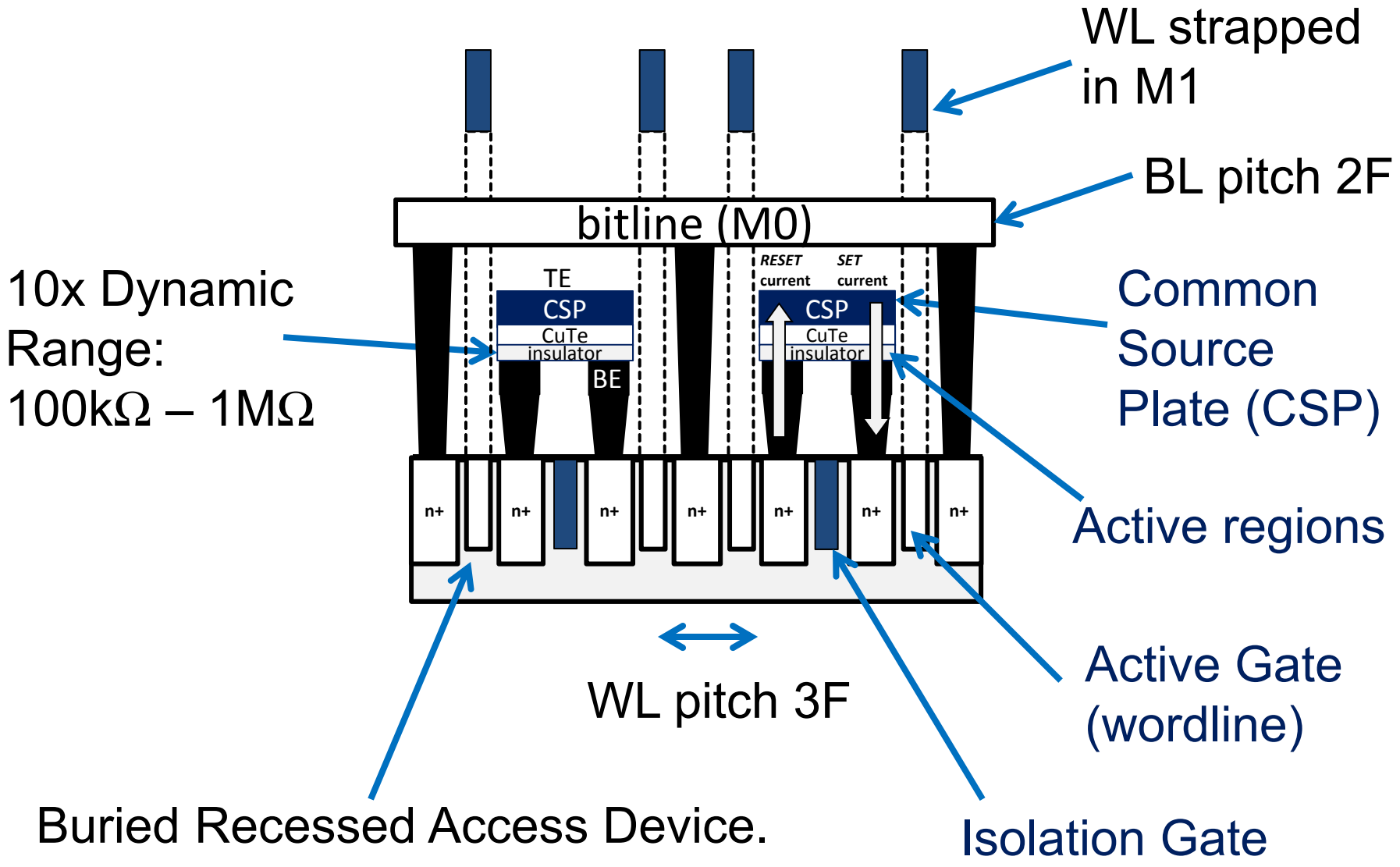
SCM-type Memory Values

- *Atomic Writes avoids supercap in critical systems*
- *Higher Endurance off-loads NAND Flash cycles*
- *Cache SSD for Performance*

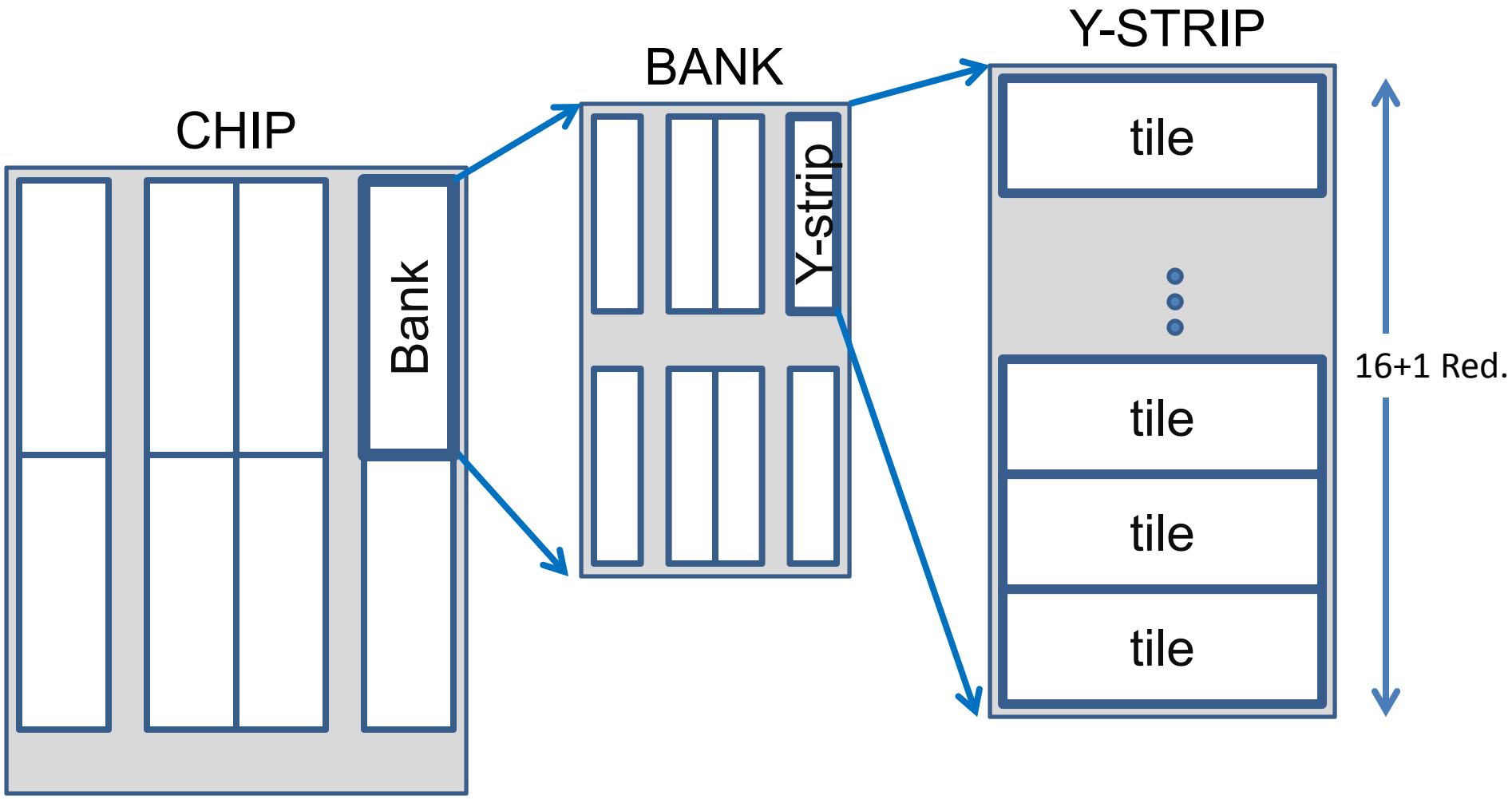
Introduction: Recent Accomplishments in ReRAM



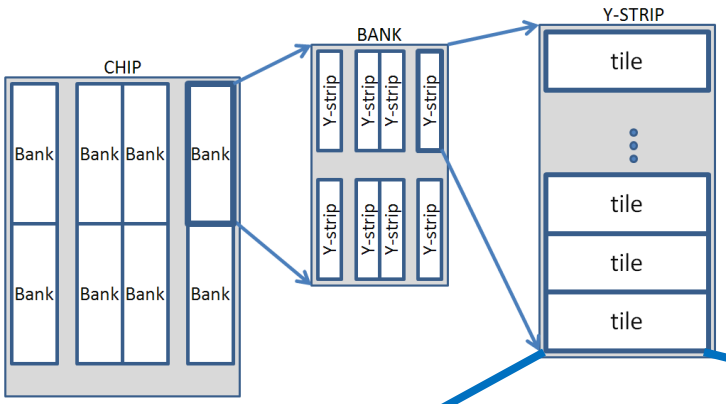
ReRAM Cell Diagram



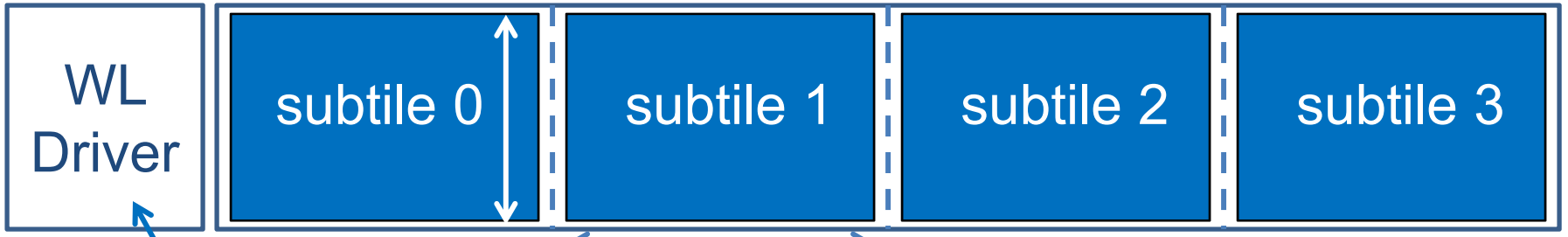
Array Architecture (1)



Array Architecture (2)



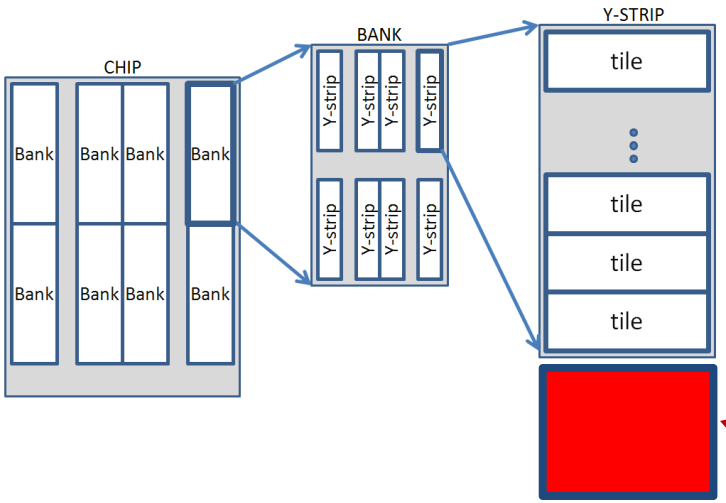
Subtile Defined by
Common Source Plate (CSP) Segment



One WL decoder shared among 4 subtiles

WL tap points between subtiles

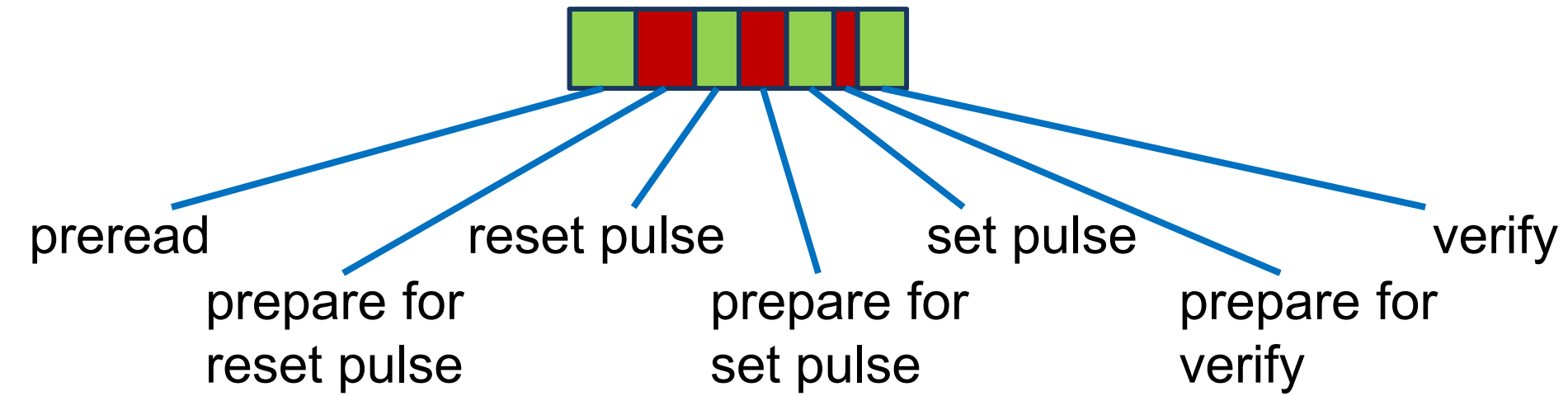
Array Architecture: Conventional



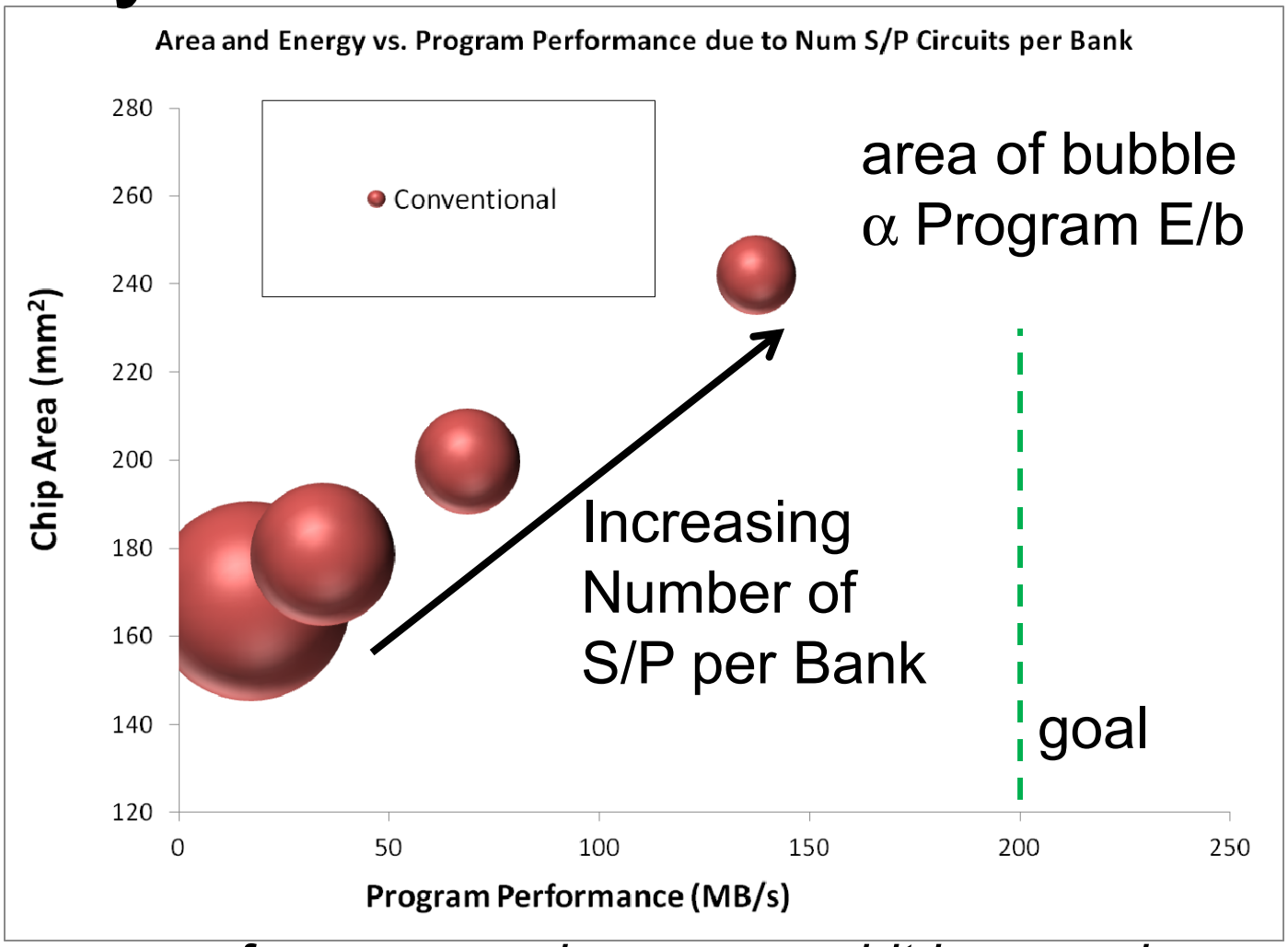
Program Efficiency
= Σ (pulse+verify)/Total
= 57%

all Sense/Program circuits (S/P) of a page

Program Algorithm

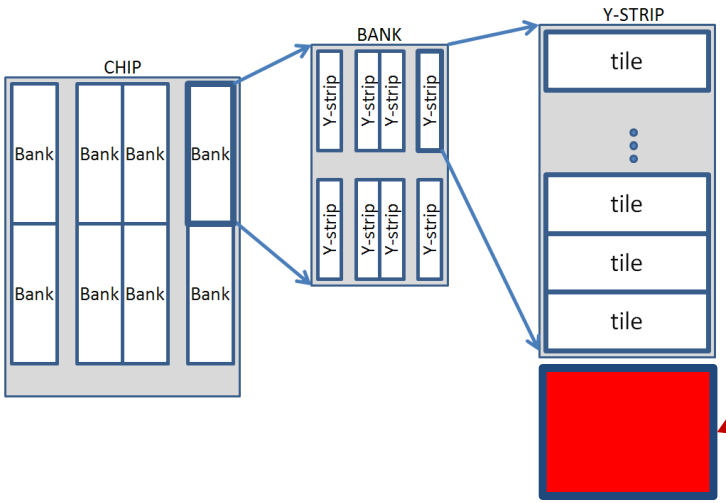


Array Architecture: Conventional



- *Program performance and energy per bit improve by increasing S/P per bank*
- *But Complex S/P Circuit → increasing number is expensive.*

Array Architecture: x4 Nibbling

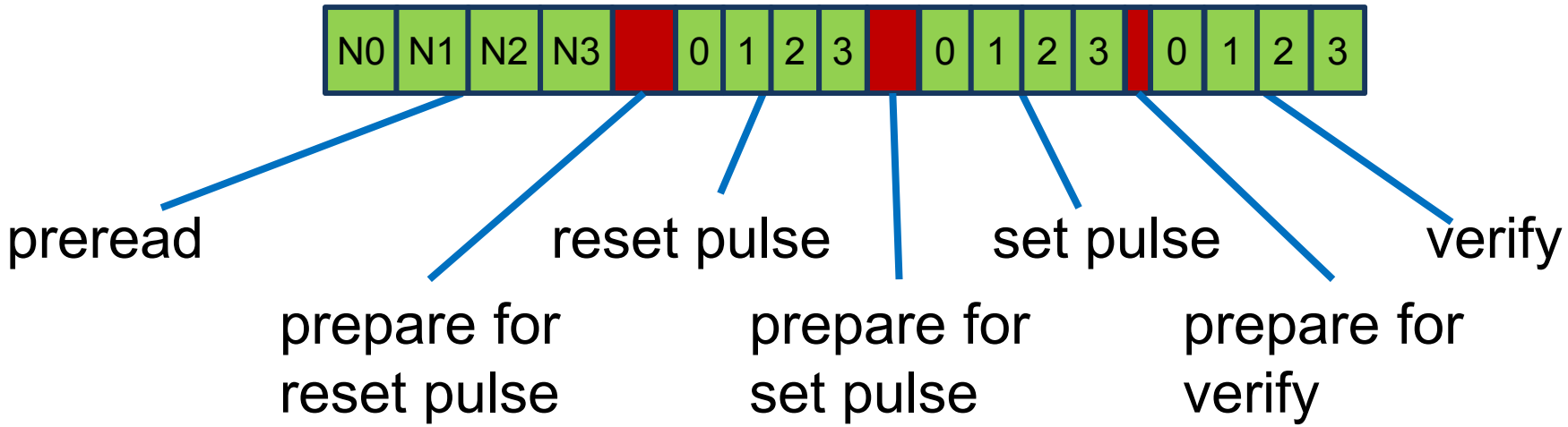


Serial Access of 4 different Y-addresses within page improves program efficiency

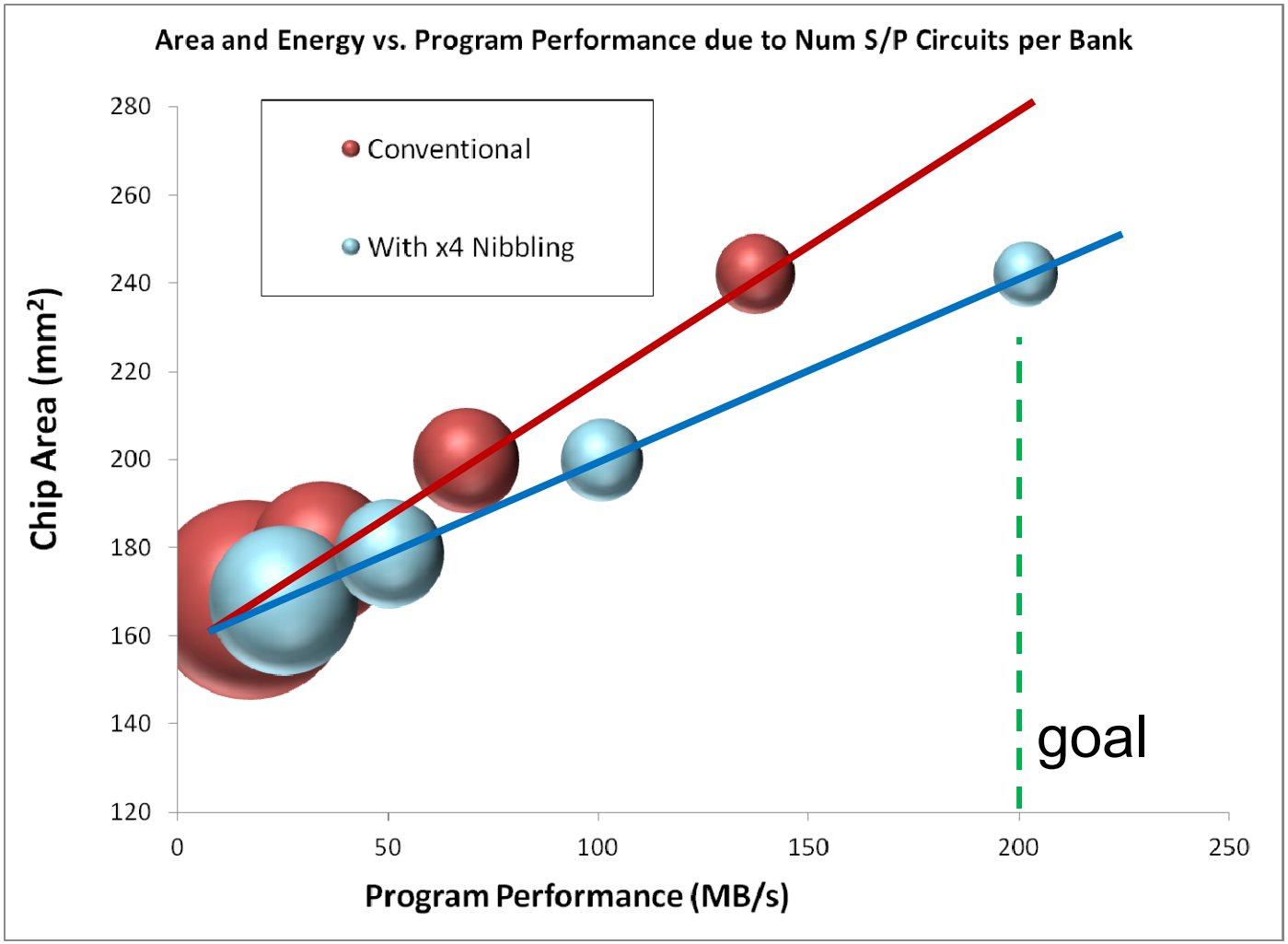
¼ of S/P of a page

Program Efficiency
= Σ (pulse+verify)/Total
= 84%

Program Algorithm

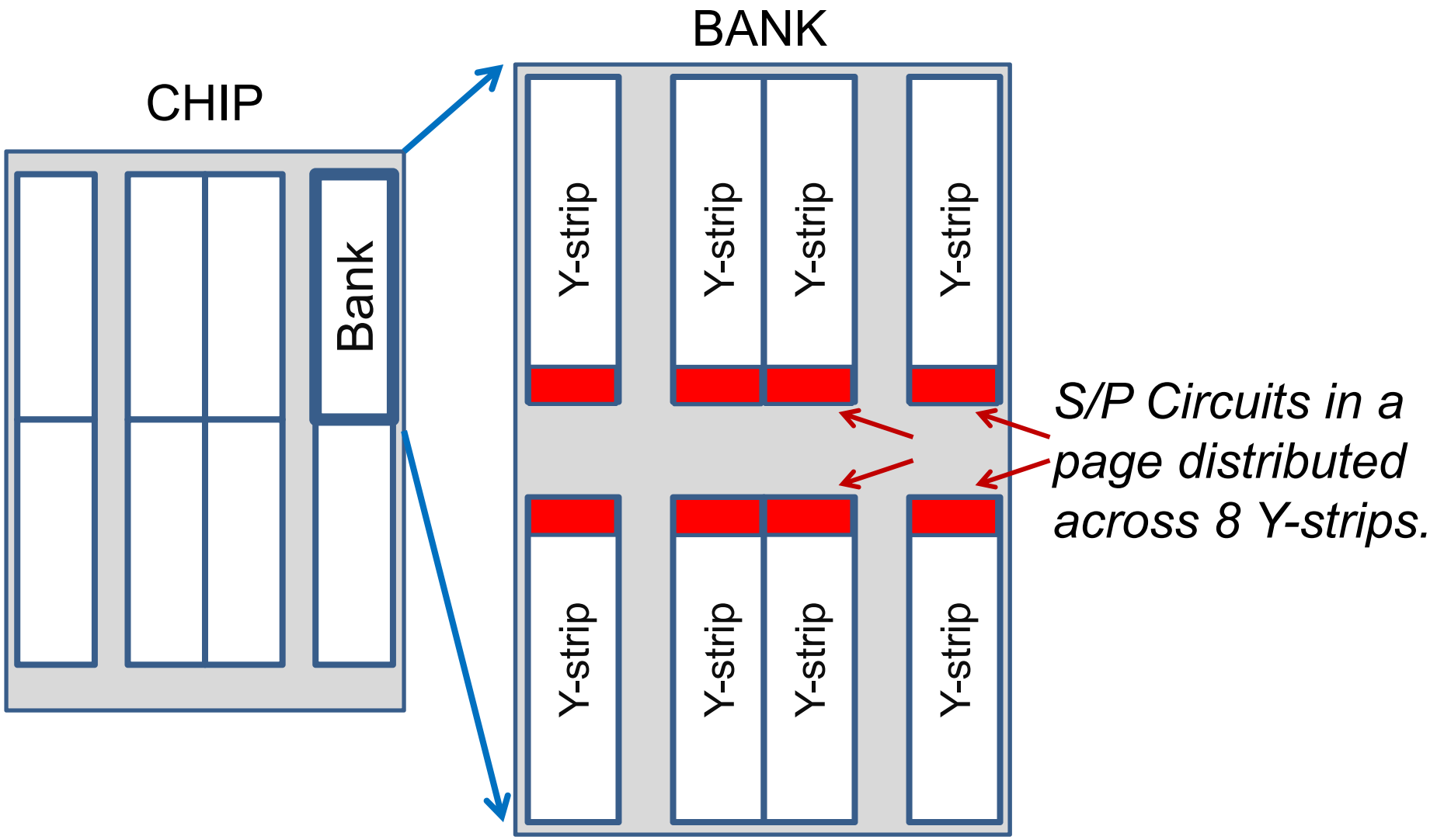


Array Architecture: x4 Nibbling

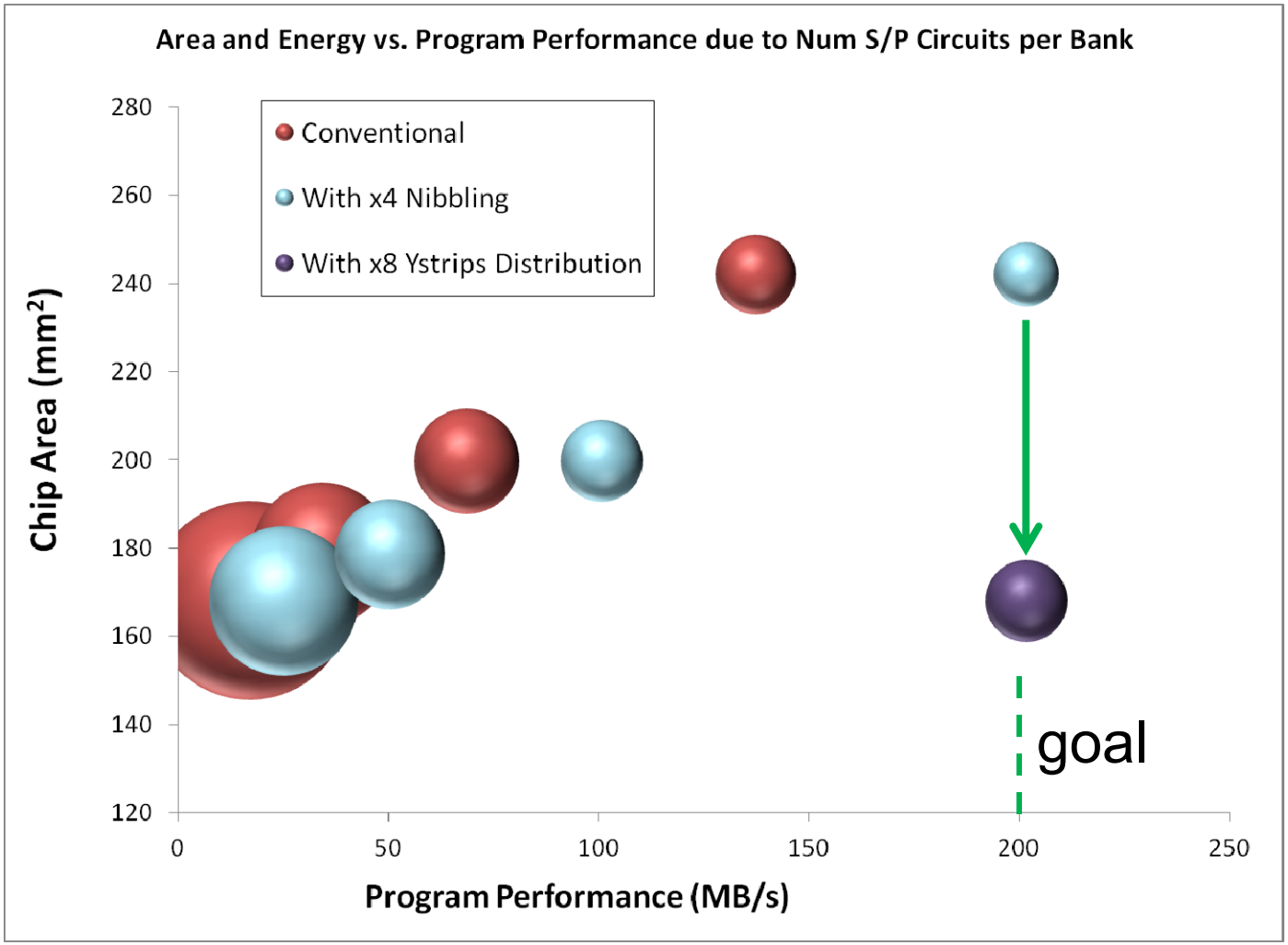


- *Improved die area for same program performance.*
- *Improved program performance and energy efficiency for same die area -- but still a very large die*

Array Architecture: Distributed Page

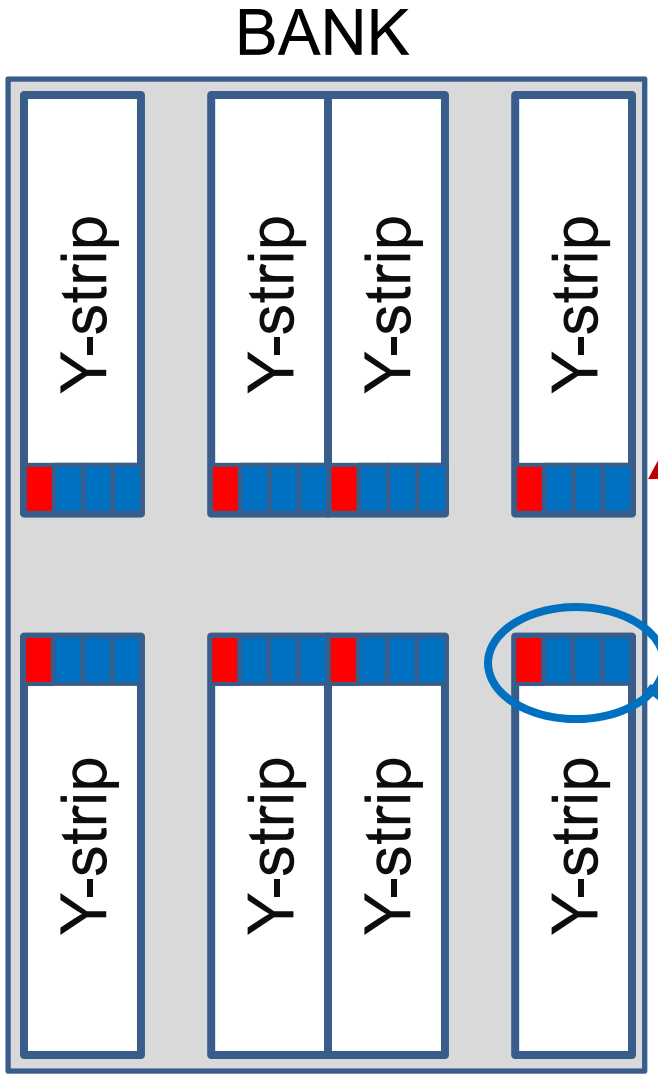


Array Architecture: Distributed Page



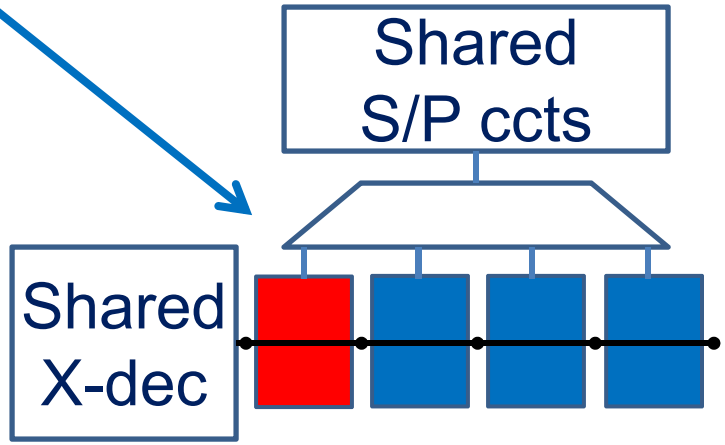
- *Die area reduction by sharing S/P circuits across Y-strips.*
- *Penalty paid in Energy per Bit due to 8x Common Source Plate (CSP) Charging.*

Array Architecture: Subtiles

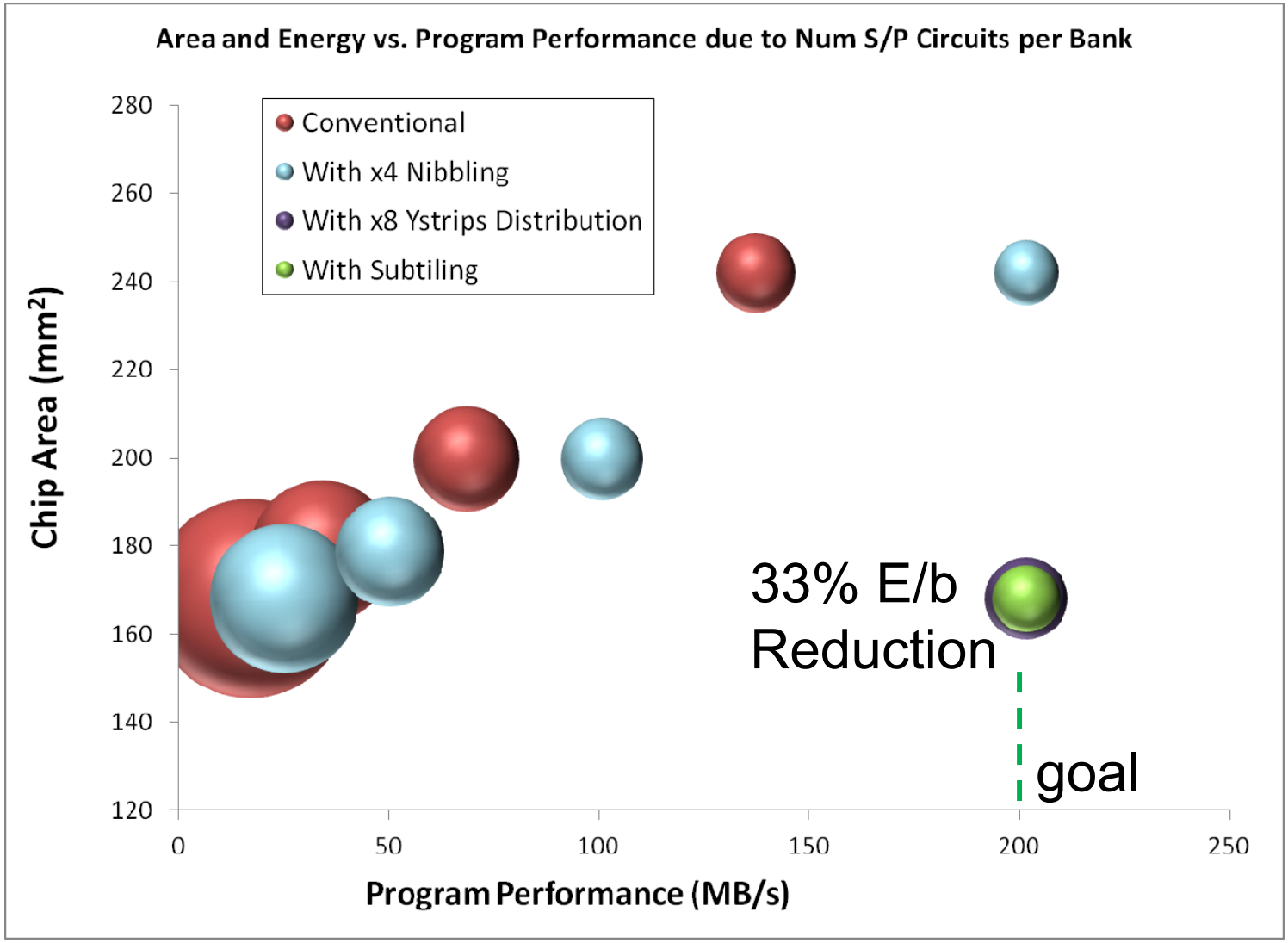


Split CSP into 4 segments "subtiles" within a tile. Reduces CSP Power

Activate one subtile in each Y-strip.

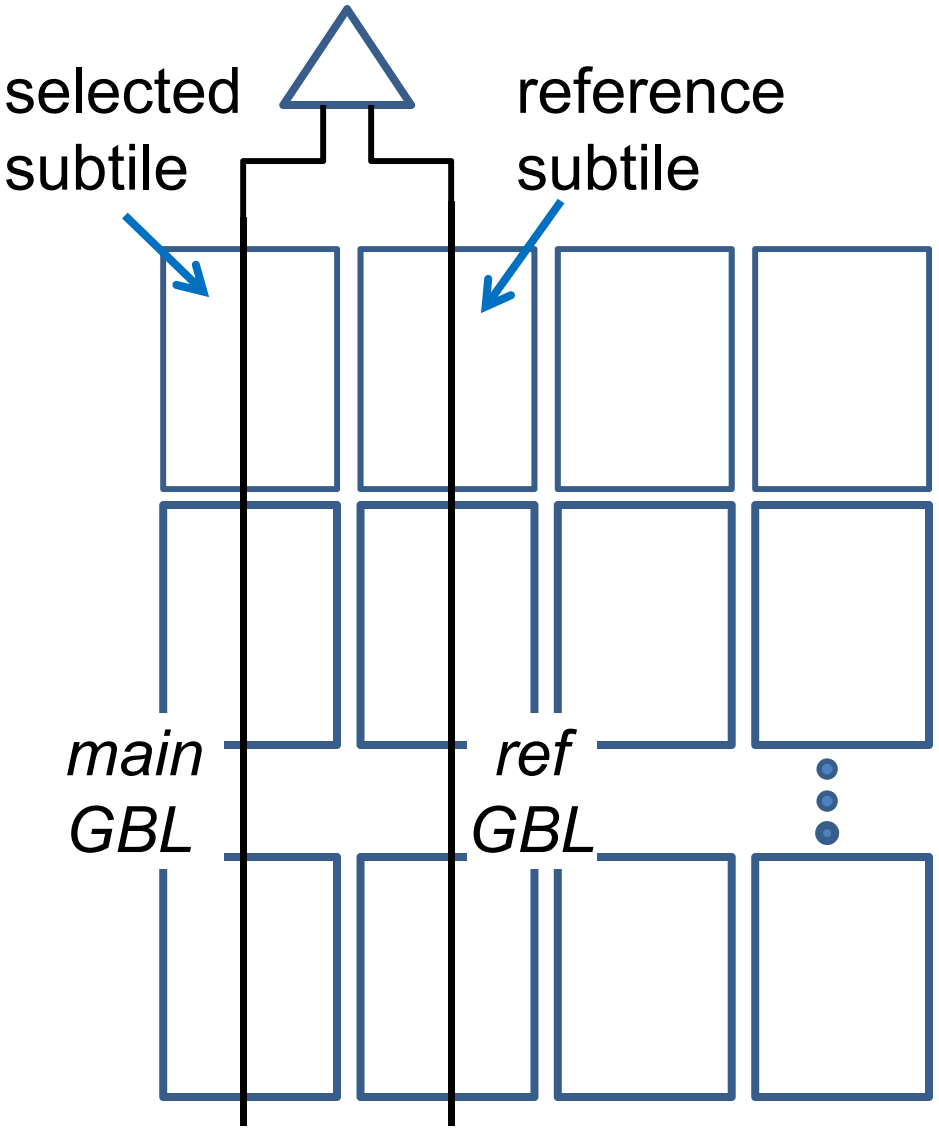


Array Architecture: Subtiles

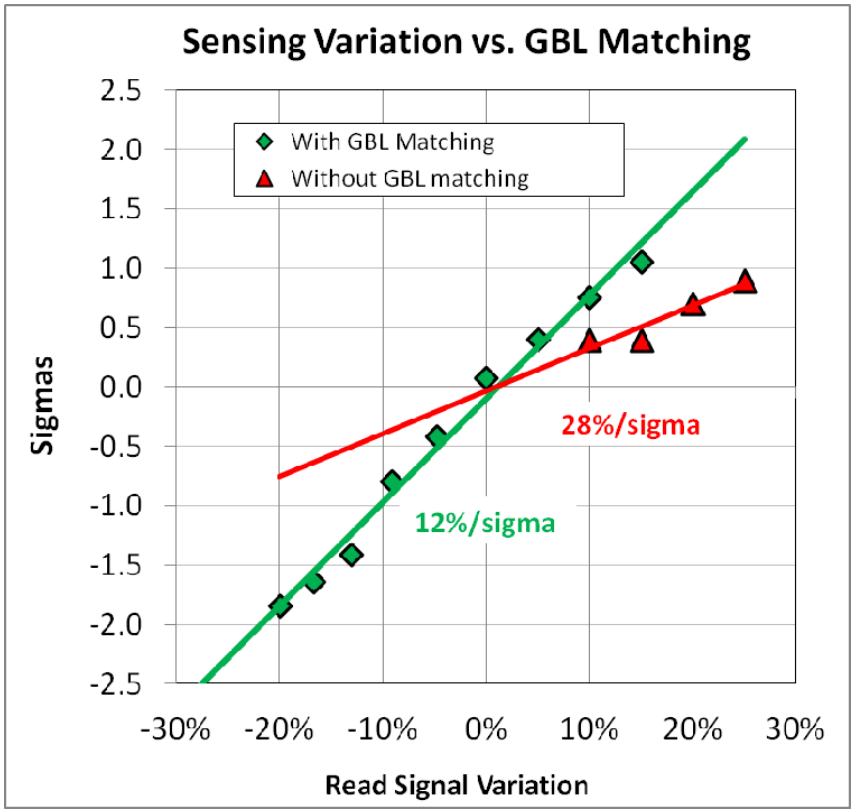


- *33% Reduction in Energy per Bit due to Smaller CSP*

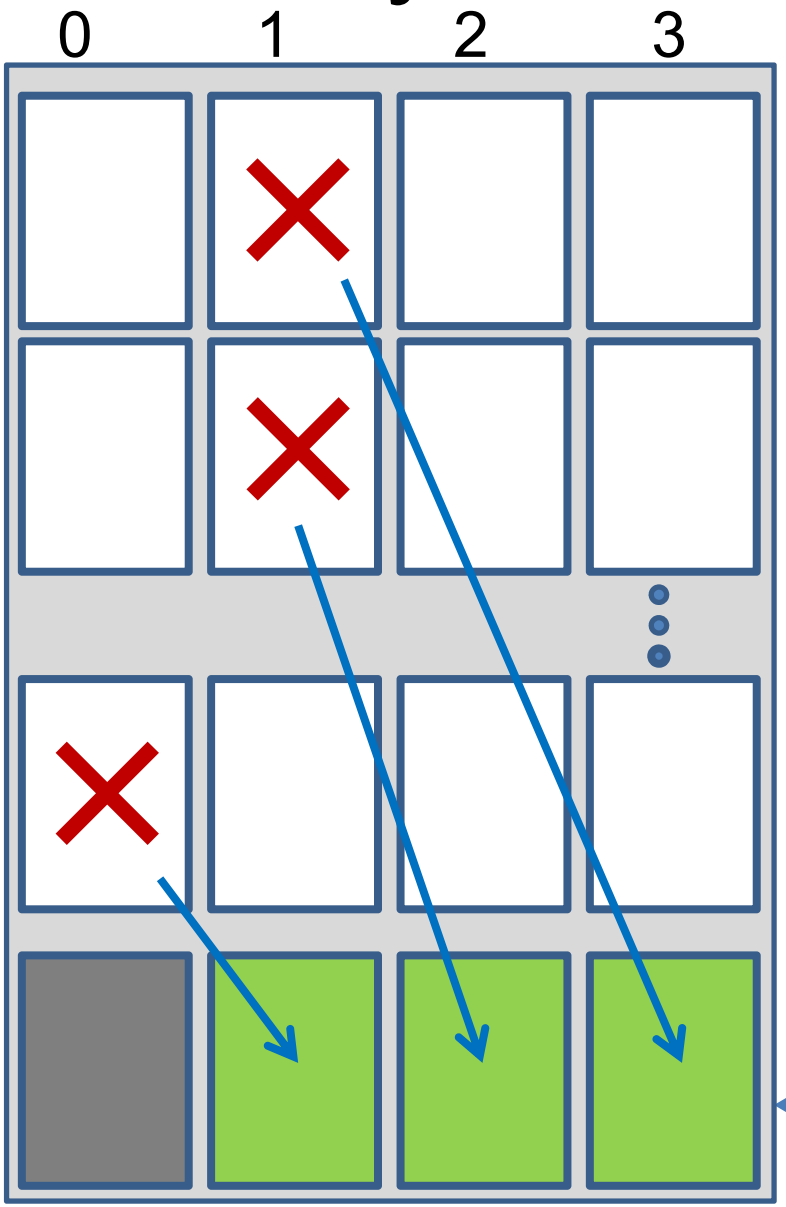
Array Architecture: Subtiles



Nearby unused subtiles also provide convenient reference Global Bitline (GBL) for good noise rejection and transient matching.



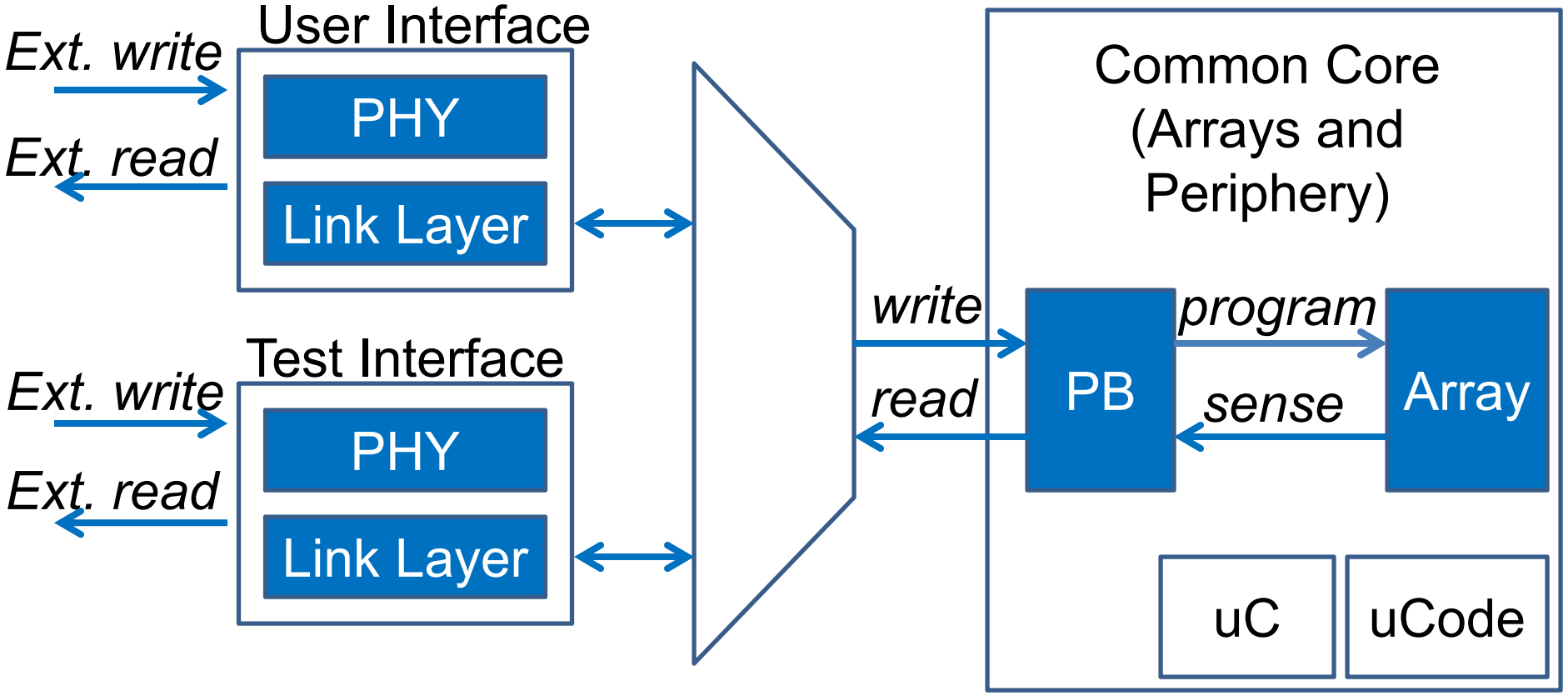
Array Architecture: Subtiles



Subtiles also allows more effective subtile redundancy. e.g. better defense against shorted CSP.

Up to 4 subtiles in a Y-strip can be cross-repaired.

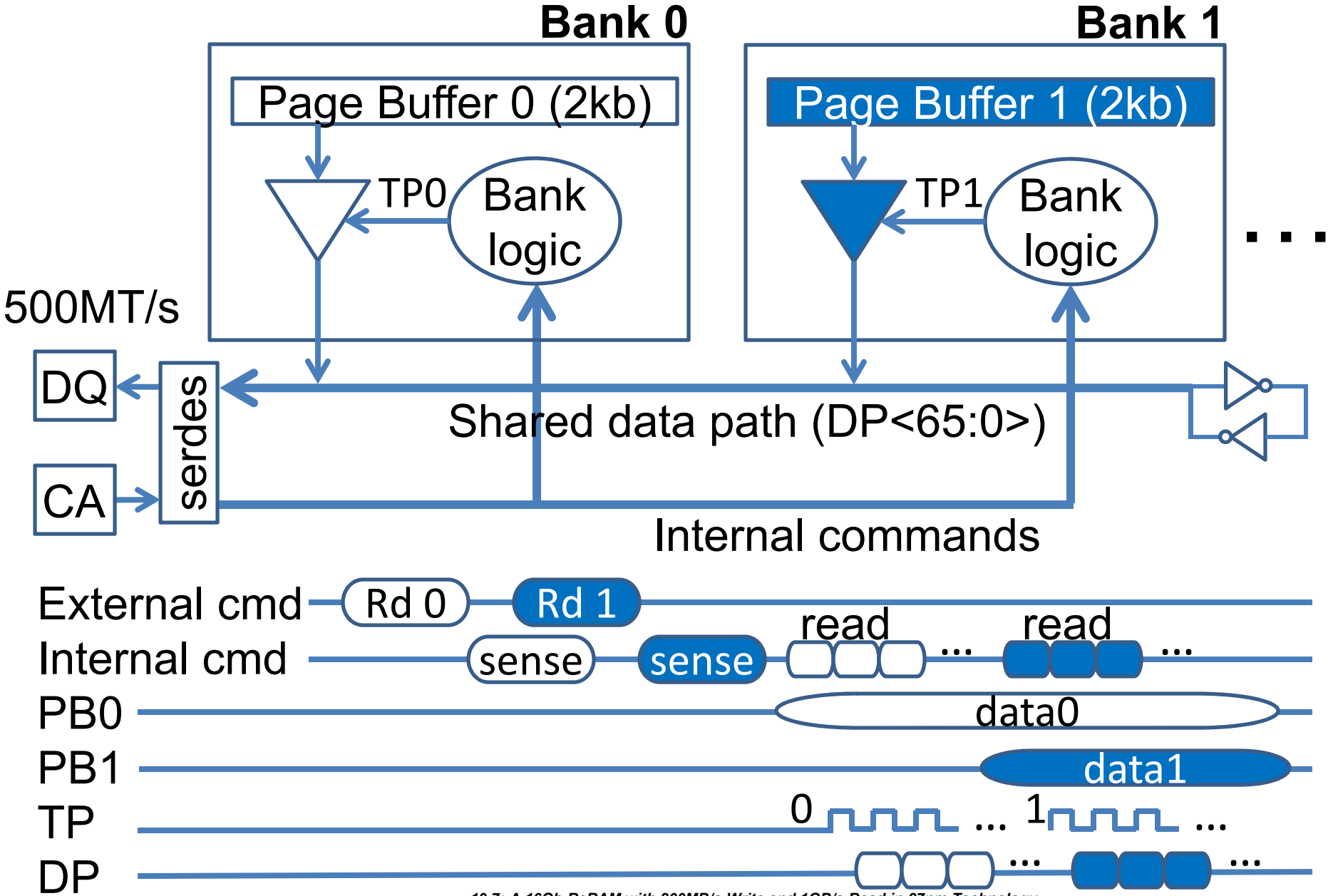
Interfaces and Data path



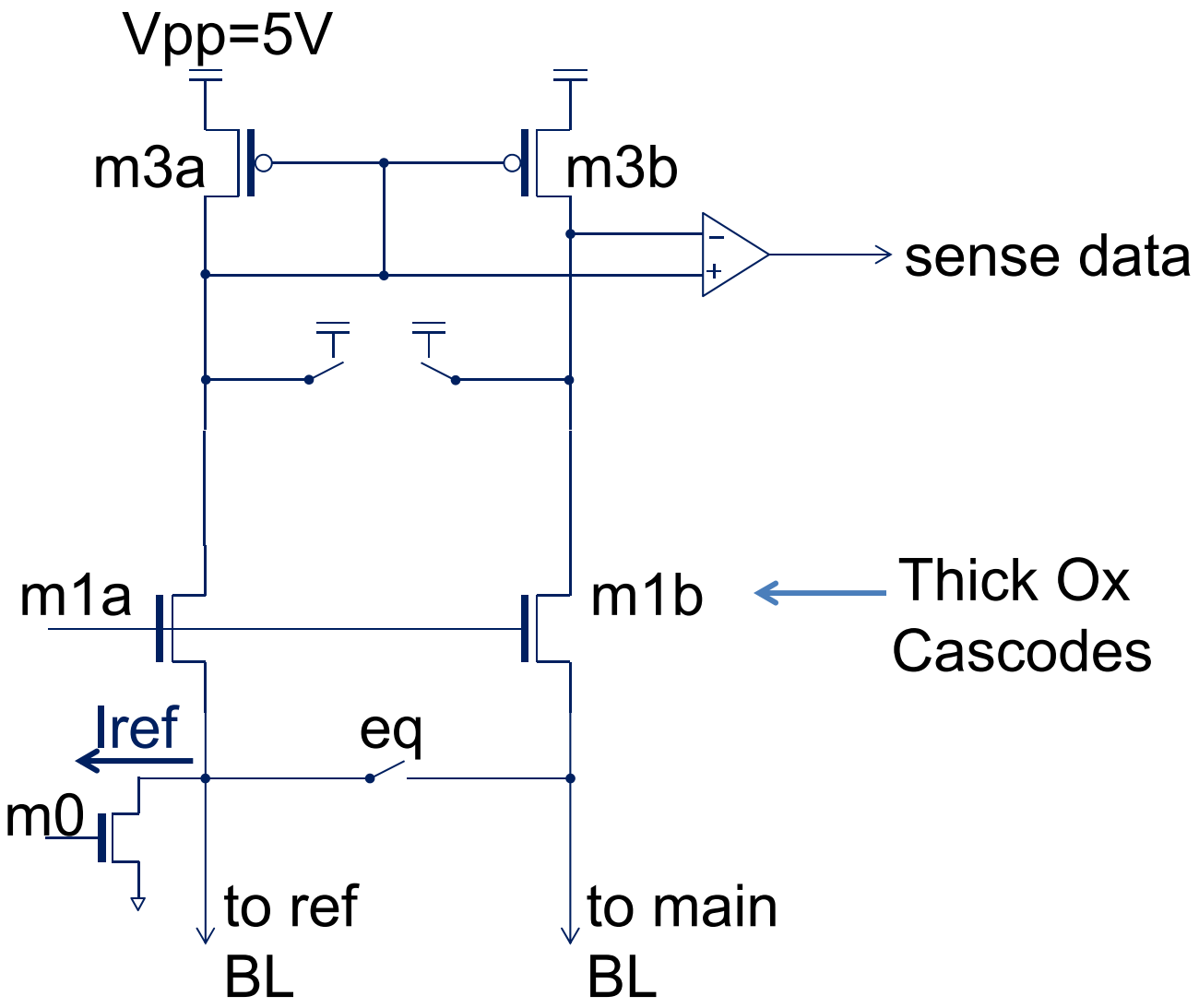
"Ext. write" ——→ 32 x write (to PB) + Program (to array)

"Ext. read" ——→ Sense (from array) + 32 x read (from PB)

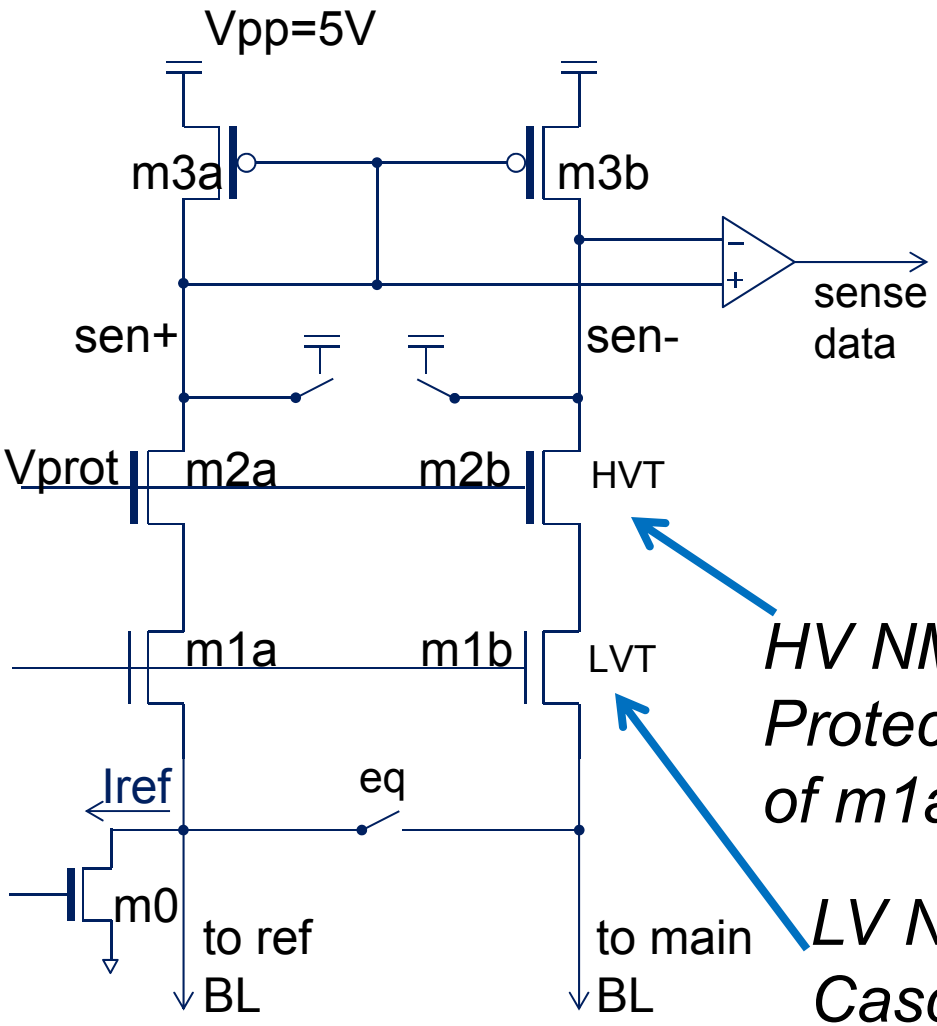
User Interface and Data Path

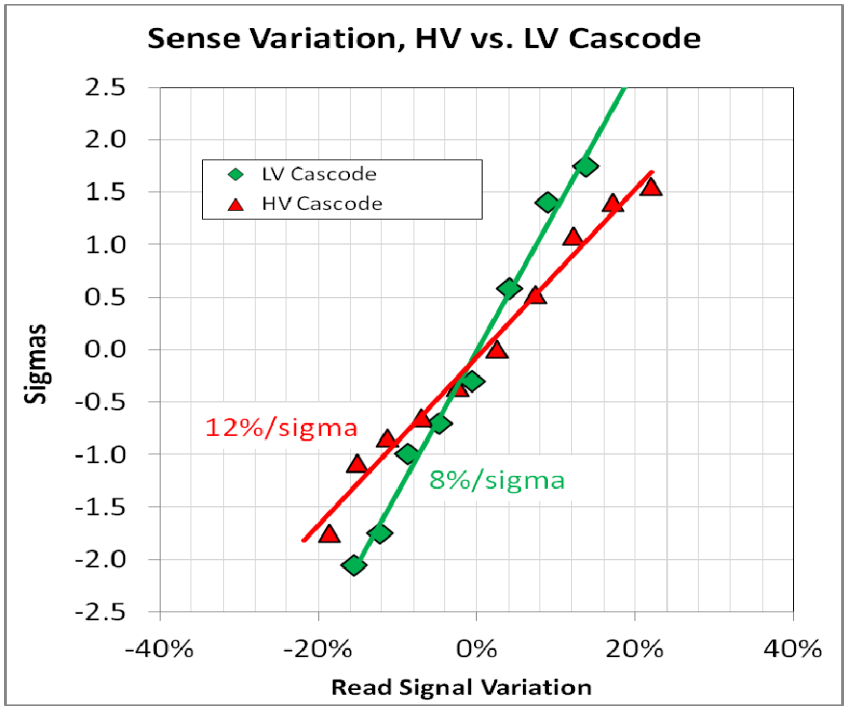


Sense Amplifier: Conventional



Sense Amplifier: Cascode Matching





HV NMOS Protection to limit Vd of m1a/b.

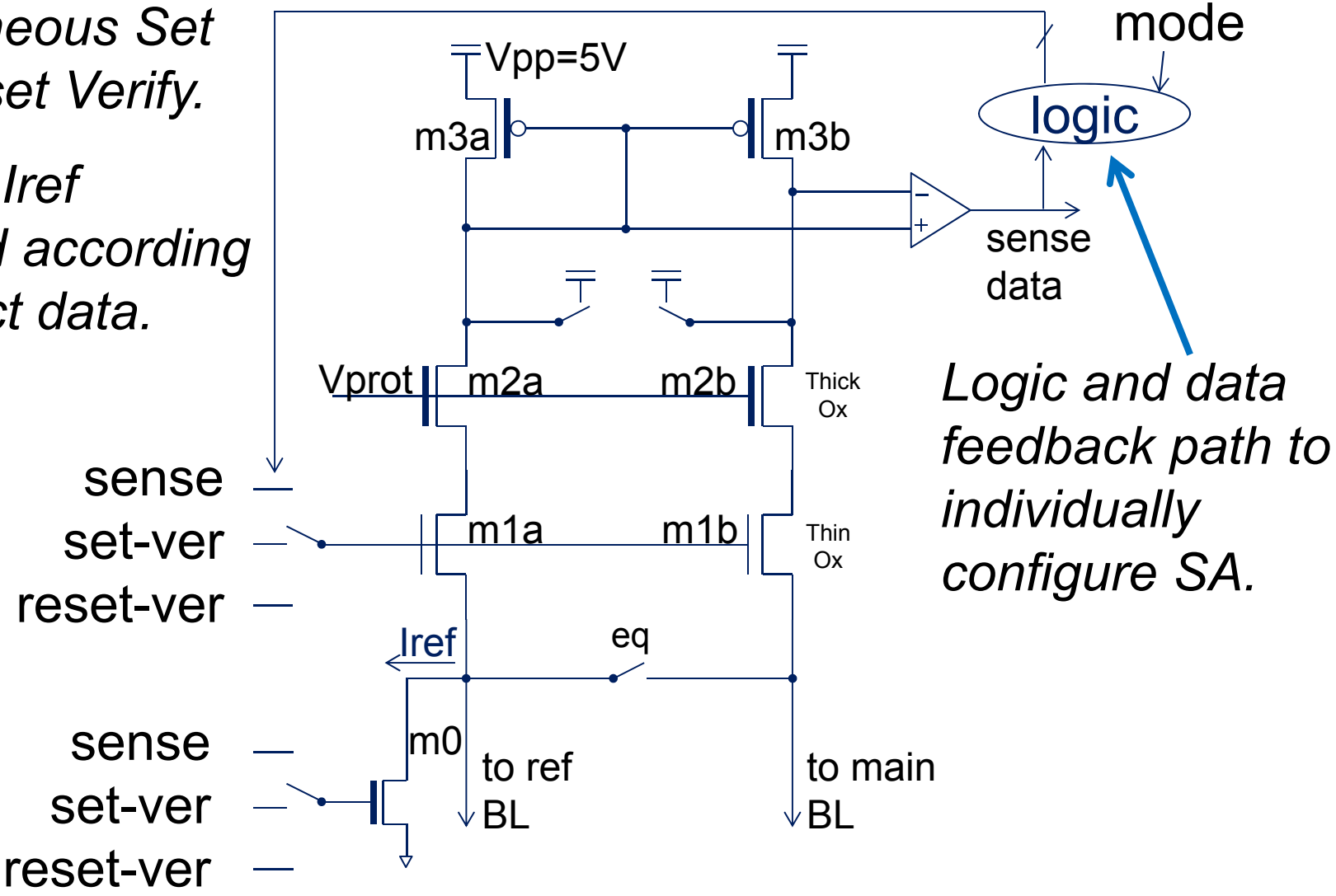
LV NMOS Cascode Pair for better matching

Also reduces cap on sen+/- nodes (m2 are smaller)

Sense Amplifier: Smart Verify

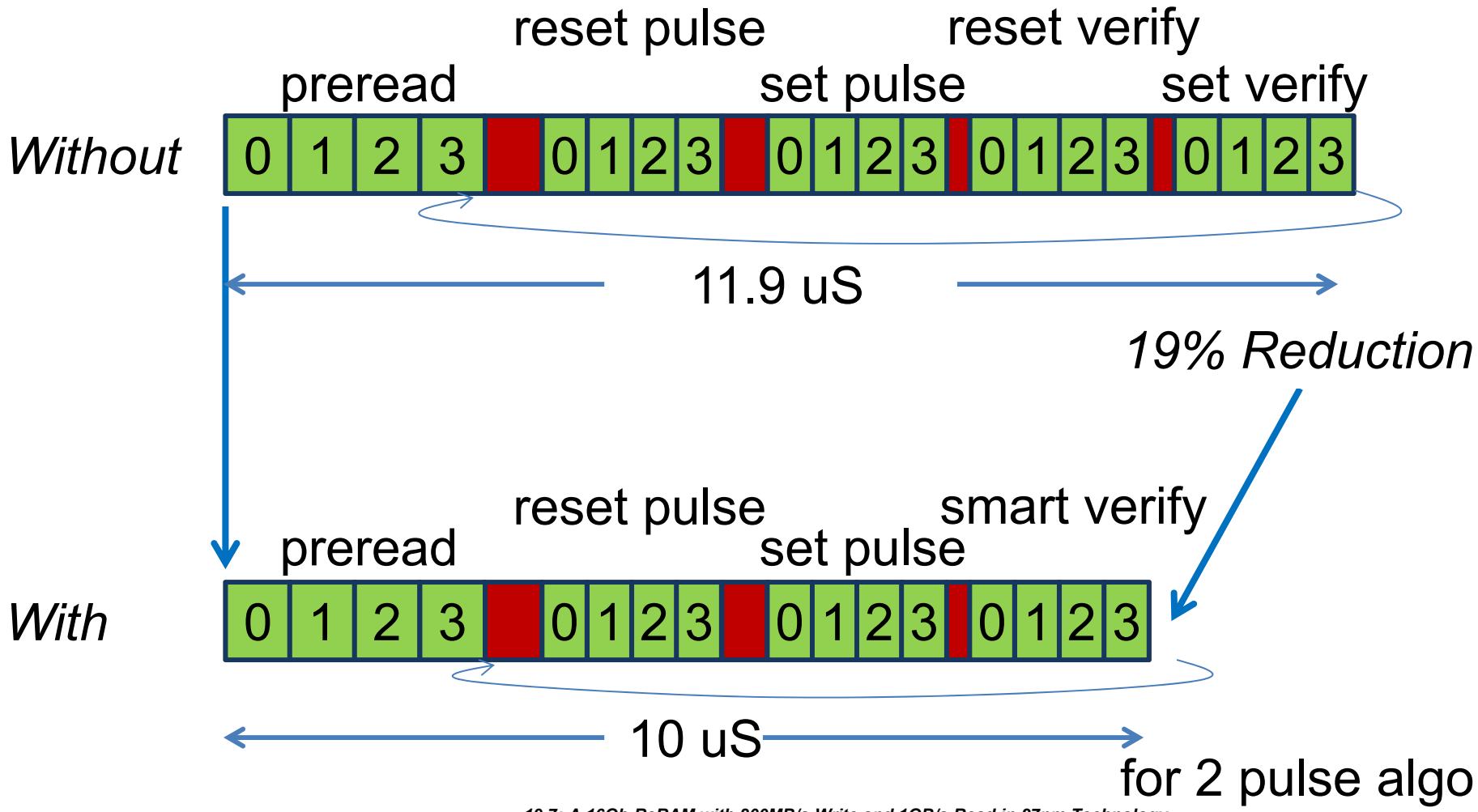
Smart Verify *allows simultaneous Set and Reset Verify.*

Vbl and Iref adjusted according to expect data.



Logic and data feedback path to individually configure SA.

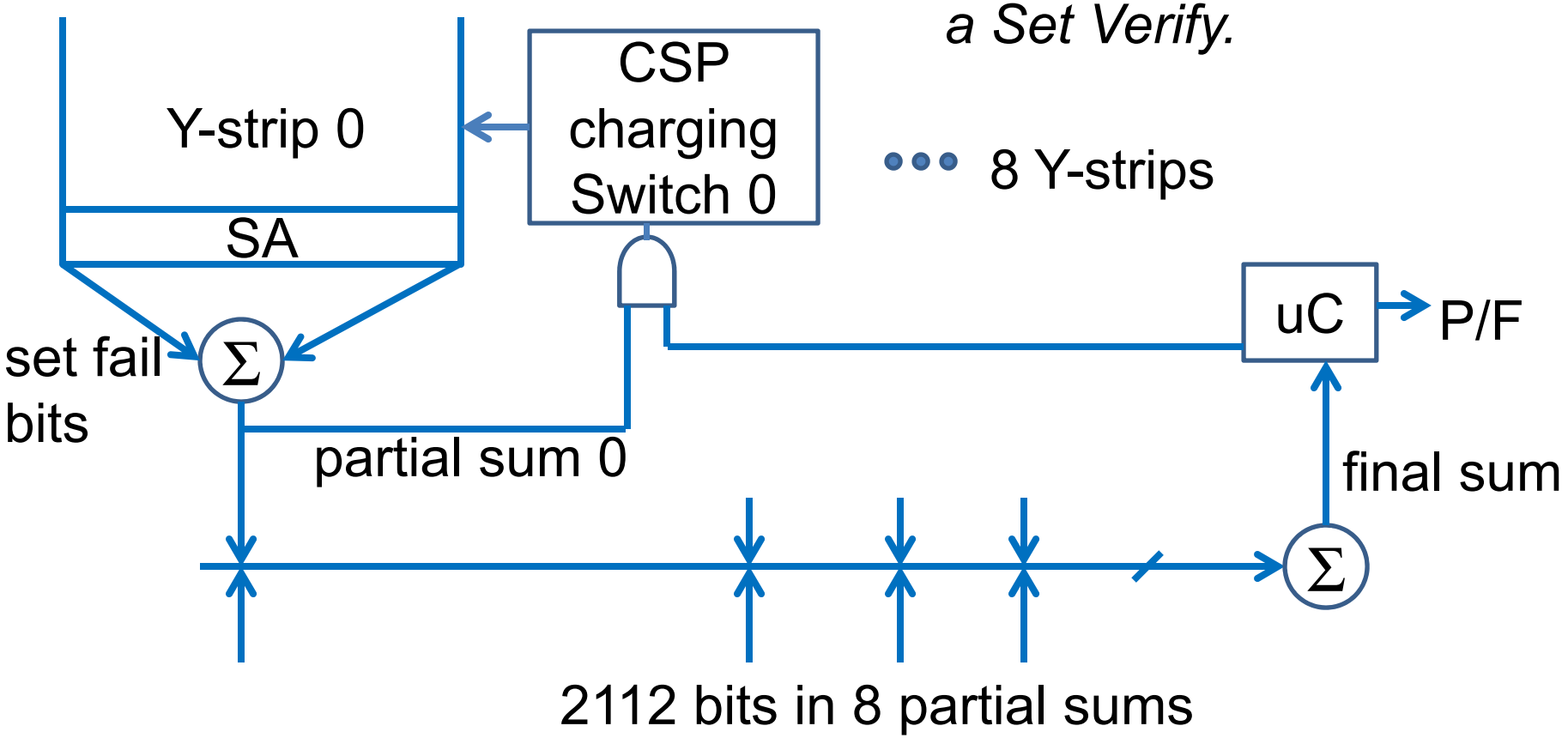
Sense Amplifier: Smart Verify Performance



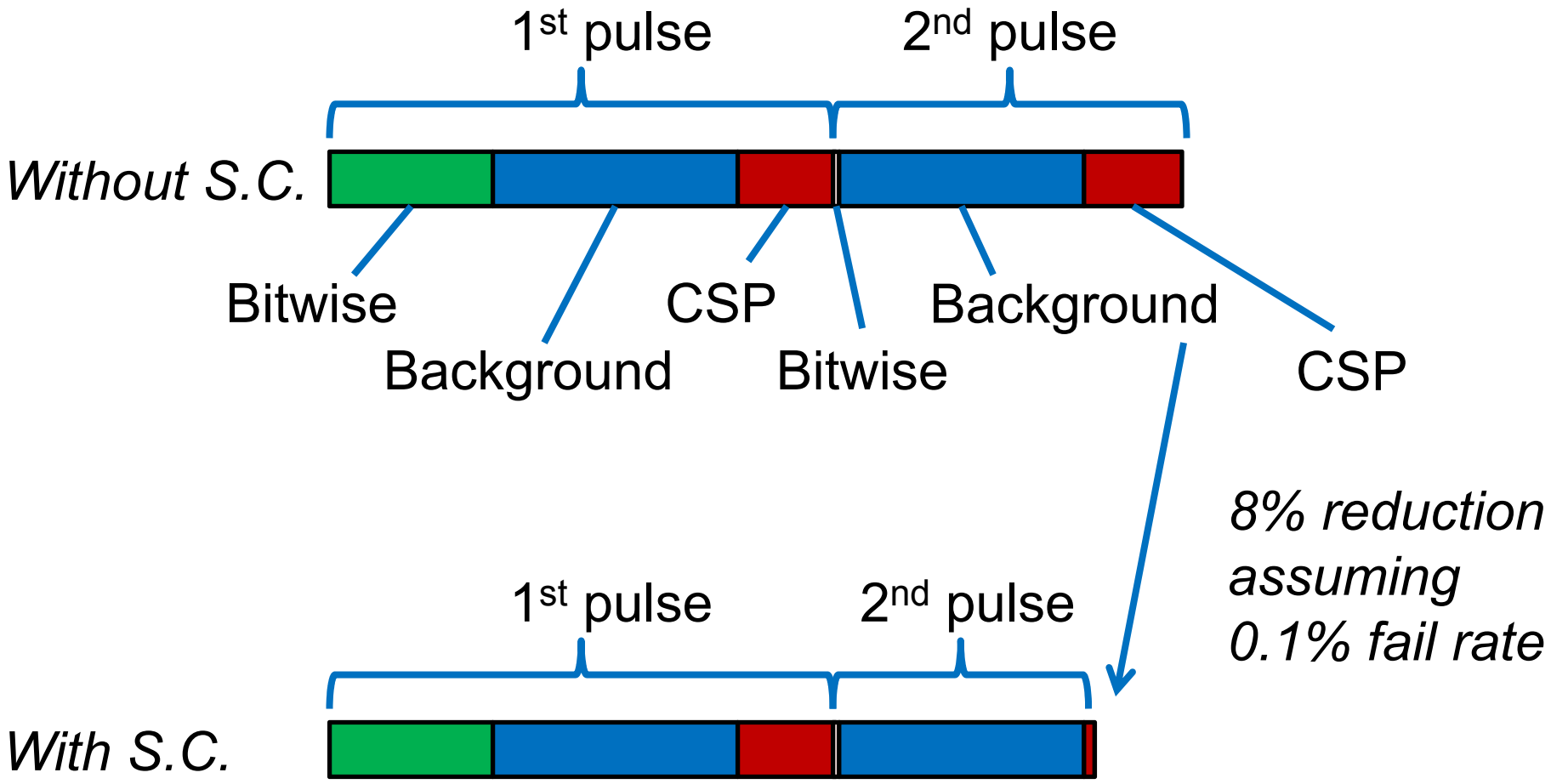
Smart Charging

Small number of bits may fail first program pulse and require a 2nd pulse.

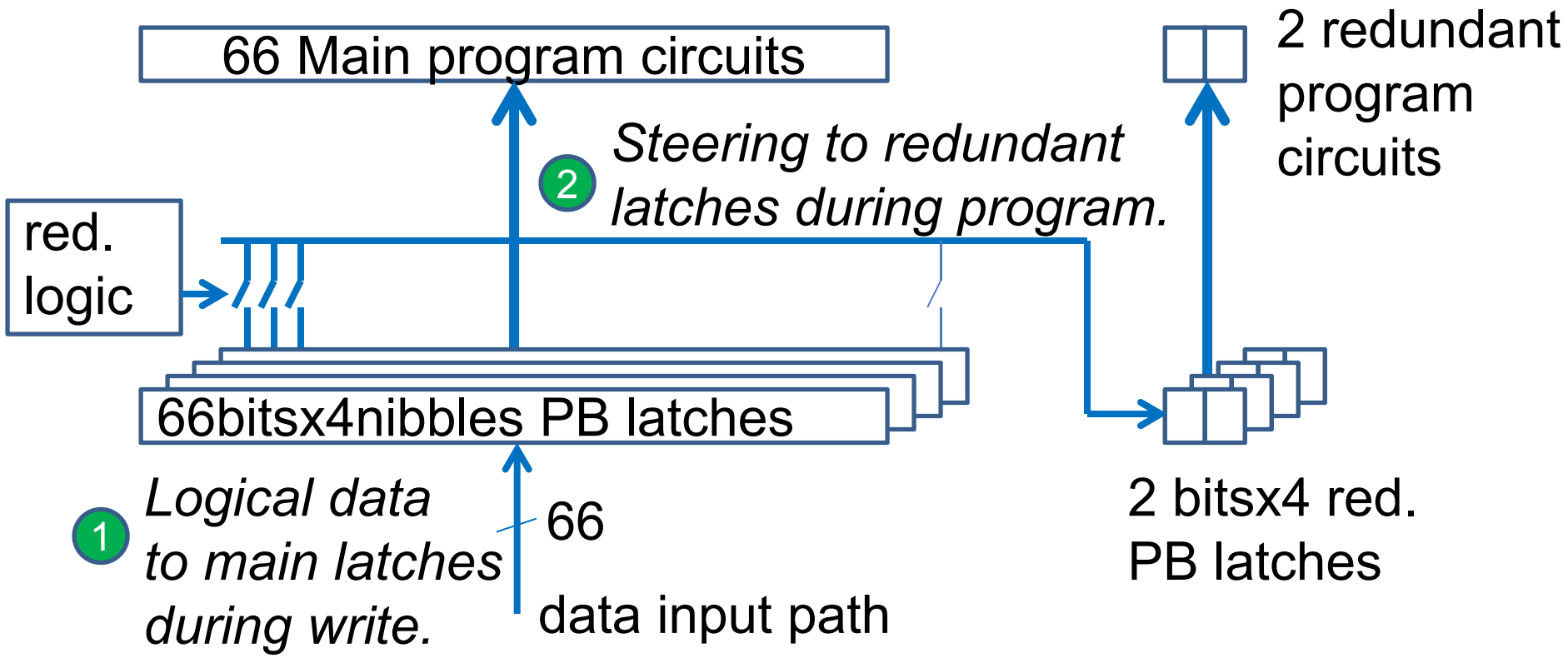
***Smart Charging** = Charge only those CSP in which cells fail a Set Verify.*



Smart Charging: E/b Savings

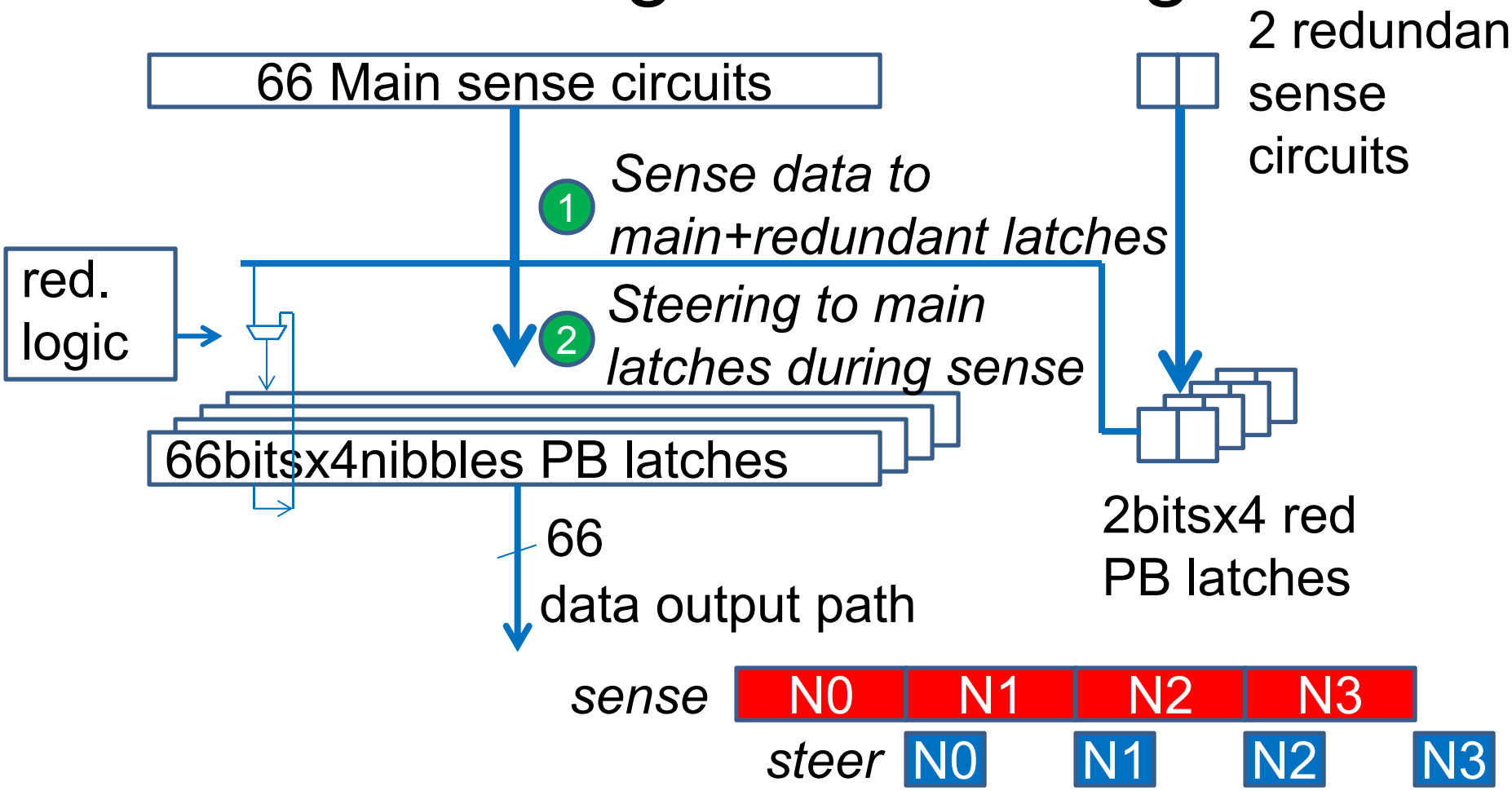


Column Redundancy: Writing and Programming



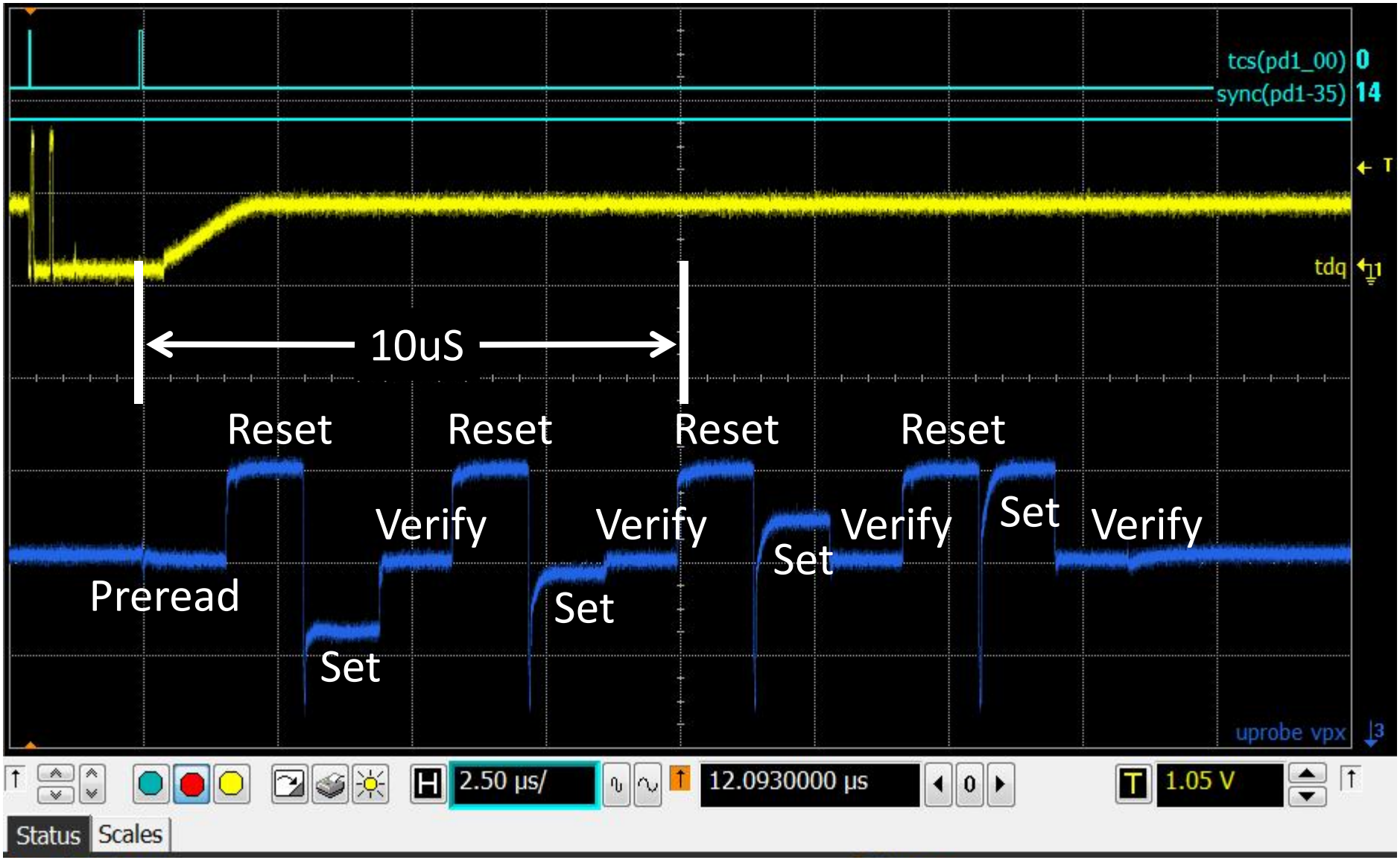
- *Steering during longer program algo relieves speed path.*
- *Steering at same time HV nodes ramp.*
- *Original data kept in PB allows it to be used at other addresses.*

Column Redundancy: Sensing and Reading



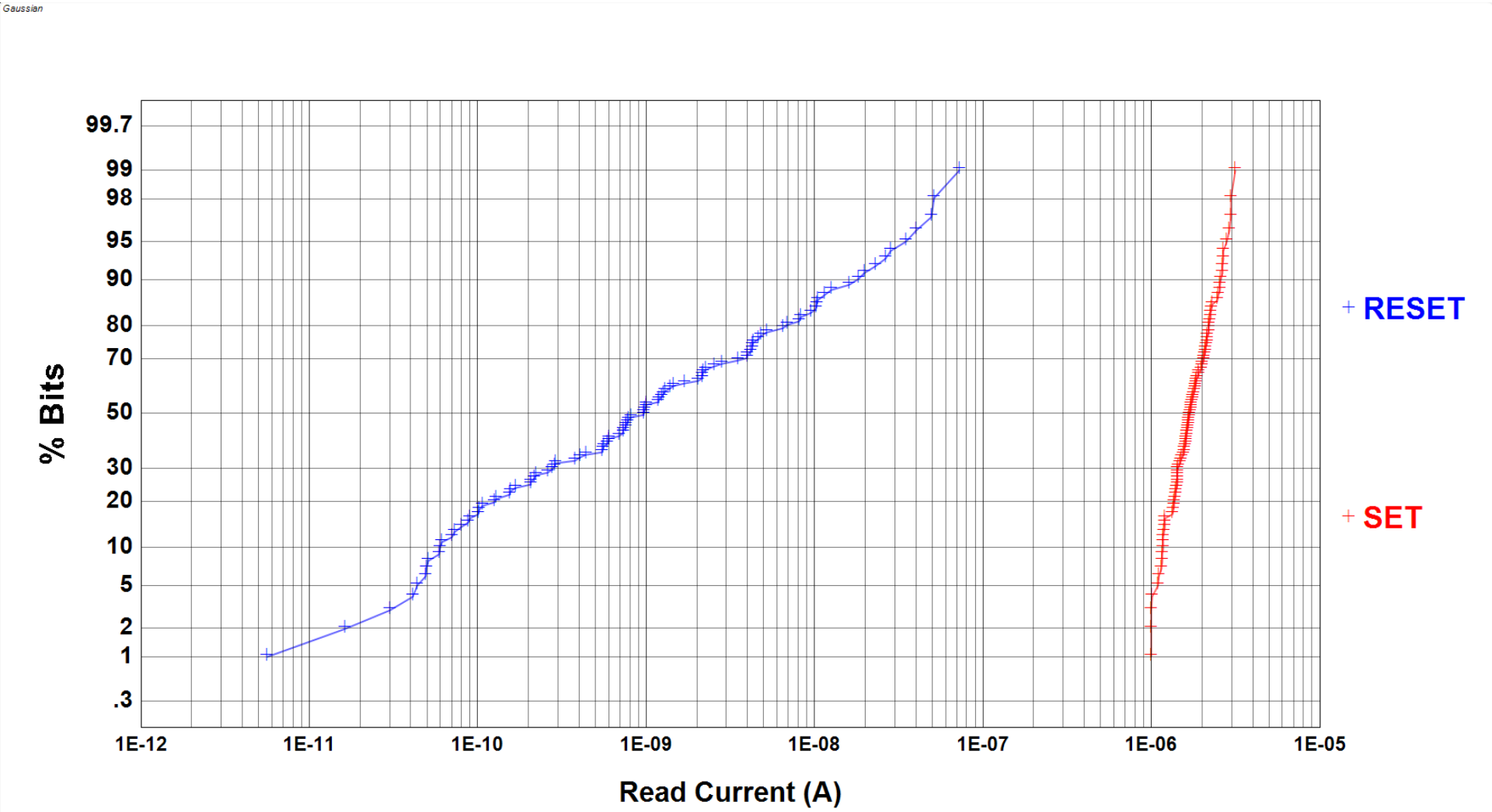
- *Steering during longer sense algo relieves speed path.*
- *Pipelining steering while sensing next nibble.*

Program Algorithm Sequence



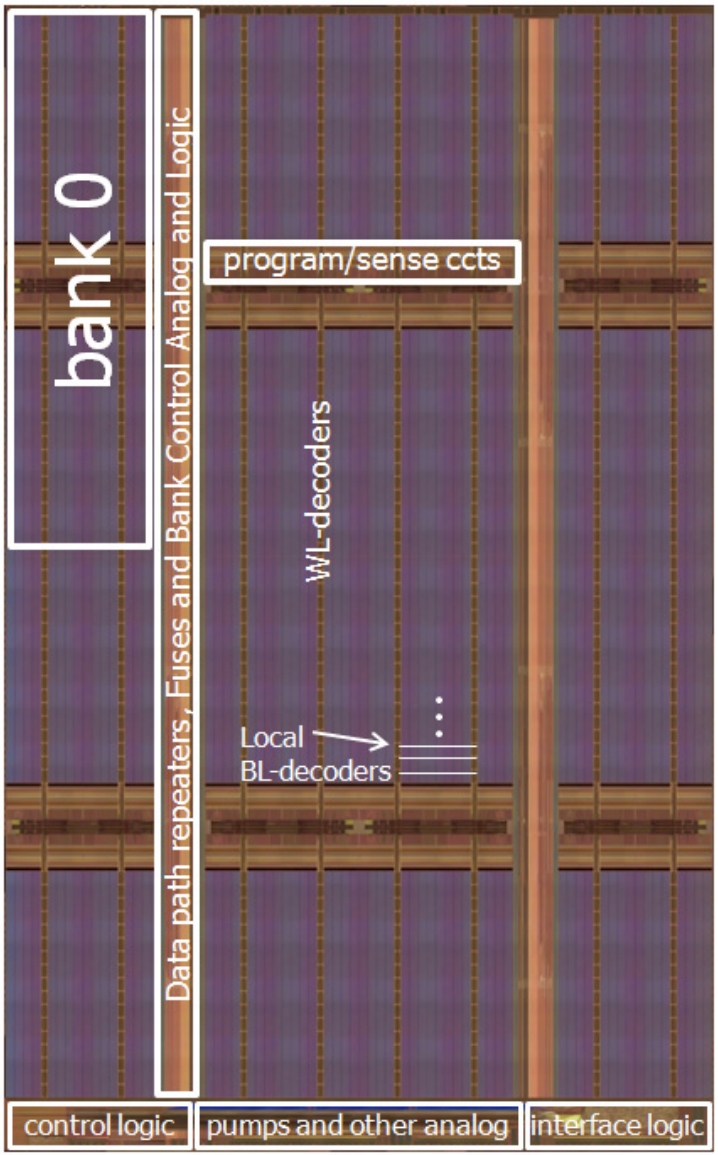
Si Measurements: IO

Set and Reset Distribution



Summary Table and Micrograph

Summary Table		
Density		16 Gb
Tech node (nm)		27
Cell Size (nm^2)		4374 ($6F^2$)
Die Size (mm^2)		168
Selector		Buried WL MOS selector
Read Performance	BW (MB/s)	1000
	Latency (uS)	2
Write Performance	BW (MB/s)	200
	Latency (uS)	10



Conclusion

- A 16Gb ReRAM with 200MB/s write and 1GB/s read throughput has been presented.
- Power, performance and die size were optimized through nibbling, a distributed page and subtile architecture.
- The user interface supports high core bandwidth with a flexible Common Core design approach.
- The sense amplifier supports Smart Verify to improve program throughput.
- Smart Charging offers a further reduction in program energy.
- Column redundancy during algorithms and with pipelining also helps to improve performance.